Conclusion of Part 3 - Classification

In this Part 3 you learned about 7 classification models. Like for Part 2 - Regression, that's quite a lot so you might be asking yourself the same questions as before:

    1. What are the pros and cons of each model ?

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

    3. How can I improve each of these models ?

Again, let's answer each of these questions one by one:

1. What are the pros and cons of each model ?

Please find here a cheat-sheet that gives you all the pros and the cons of each classification model.

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

Same as for regression models, you first need to figure out whether your problem is linear or non linear. You will learn how to do that in Part 10 - Model Selection. Then:

If your problem is linear, you should go for Logistic Regression or SVM.

If your problem is non linear, you should go for K-NN, Naive Bayes, Decision Tree or Random Forest.

Then which one should you choose in each case ? You will learn that in Part 10 - Model Selection with k-Fold Cross Validation.

Then from a business point of view, you would rather use:

- Logistic Regression or Naive 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. Eventually that allows you to target your marketing campaigns. And of course for this type of business problem, you should use Logistic Regression if your problem is linear, and Naive Bayes if your problem is non linear.

- SVM when you want to predict to which segment your customers belong to. Segments can be any kind of segments, for example some market segments you identified earlier with clustering.

- Decision Tree when you want to have clear interpretation of your model results,

- Random Forest when you are just looking for high performance with less need for interpretation.

3. How can I improve each of these models ?

Same answer as in Part 2:

In Part 10 - Model Selection, you will find the second section dedicated to Parameter Tuning, that will allow you to improve the performance of your models, by tuning them. You probably already noticed that each model is composed of two types of parameters:

- the parameters that are learnt, for example the coefficients in Linear Regression,

- the hyperparameters.

The hyperparameters are the parameters that are not learnt and that are fixed values inside the model equations. For example, the regularization parameter lambda or the penalty parameter C are hyperparameters. So far we used the default value of these hyperparameters, and we haven't searched for their optimal value so that your model reaches even higher performance. Finding their optimal value is exactly what Parameter Tuning is about. So for those of you already interested in improving your model performance and doing some parameter tuning, feel free to jump directly to Part 10 - Model Selection.

Now congratulations for having completed Part 3, and let's move on the next part of the journey:

Part 4 - Clustering