1. How do you deal with an imbalanced data-set, assuming SMOTE is not an option?

Different strategies can be used to minimize the accuracy paradox derived from an imbalanced data-set. This kind of data-set will lead frequently to models trained to high accuracies that in fact just reflect the out-balanced distribution of data tilted to (usually) one specific class.

One tactic to deal with an imbalanced data-set is to select different performance metrics when accuracy is tainted by the population distribution towards a specific class, p.ex: confusion matrix, precision, recall, f-score, Cohen's kappa and/or ROC Curves.

Collecting more data may also provide a more balanced perspective on the data-set classes. If totally impossible, one may instead resample the data-set adding copies of under-represented classes (over-sampling) or removing instances of the over-represented class (under-sampling). Another strategy may involve decomposing the larger class into a small number of other classes.

Some algorithms are known to perform well on imbalanced data-sets - like decision trees - while others introduce penalties for making classification mistakes - penalized-SVM and penalized-LDA. One can also try different perspectives and apply other fields of study that may prove more adequate like anomaly detection or change detection.

2. Explain in your words what model calibration is.

Model calibration is the systematic adjustment of the parameters that describe a model so it represents a process or future events within an expected degree of uncertainty.

3. When is explainability important?

Explainability is important when the decisions of automated systems have relevant impact on the individual lives of those affected by them, either from incorrect diagnosis - in health care, p.ex - or loss of opportunities - in credit attribution or hiring processes.

Explainability is also important when exist regulatory enforcements like the General Data Protection Regulation (GDPR) in Europe. These enforcements probe decision making regarding business impact, regulatory compliance, technical approach, and even ethical values. AI explainability is part of data scientists responsibility to govern outcomes from systems and processes so that any type of bias - gender, ethnicity, etc - are identified and dealt with as quality assurance.

4. How do you define the right threshold for a binary decision?

The ‘right’ threshold for a binary decision depends on the balance of precision - proportion of correct positive identifications - and recall - proportion of positives correctly identified. Decreasing the classification threshold will increase the number of false positives and decrease the number of false negatives - i.e., precision decreases and recall increases. Increasing the classification threshold leads the effectiveness of the model in the opposite direction.

The choice of threshold for a binary decision depends especially on the ‘costs’ of false positives and false negatives, respectively.

5. What’s the difference between a score and a probability in a classification model?