Classification: Initial Steps

1. Load Data

- Ensure the data is loaded correctly.
- Examine the first few rows to understand its structure and contents.

2. Train / Test Split

- Decide on the proportion for your train-test split.
- Hotel cancellations are unbalanced, you should probably stratify your split. See `stratify` parameter in sklearn's `train_test_split()`.
- After splitting, set the test set aside. Do not touch or peek into it until the best models are ready for evaluation.

3. Cross-validation

- Set up a cross-validation strategy.
- Hotel cancellations are unbalanced, you should probably stratify your Kfold splits. See sklearn's `StratifiedKFold()`.
- Choose the evaluation metrics you consider appropriate for your problem (e.g., recall and precision).

4. Baseline

- Establish a simple baseline model. This will give you an initial performance metric to beat.
- Perform cross-validation on this model and record the results.

5. Logistic Regression

- Choose a range for the number of neighbors
- Train multiple logistic regression models varying the `penalty`. Plot their performances in train and validation.
- Now vary the `C` hyperparameter. Plot their performances in train and validation.
- Now vary the `solvers`. Plot their performances in train and validation.
- Perform a randomized search or grid search of hyperparameters, setting the ranges you consider more appropriate.
- Plot the confusion matrix for all the predictions of the validation data. Check and use the `cross_val_predict()` sklearn function for obtaining the validation predictions.

6. For KNN, DT, RF, GB:

- Perform a randomized search or grid search of hyperparameters, setting the ranges you consider more appropriate.
- Plot the confusion matrix for all the predictions of the validation data.
- Check the most important features for the models that allow it.

7. Predicted Probabilities

- Check the predicted probabilities of some of your models. Sklearn's classification algorithms usually implement the `predict_proba()` function.
- Are these probabilities very spread out? Is the model usually correct (i.e., higher probabilities really correspond to 1 and lower to 0)?

8. Final Comparison

- Plot a comparison of the performances (in train and validation sets) of all your best models, including the baselines.