Off-Week Activity

Take the Spam email dataset or any other dataset of your choice and apply the following operations on it.

1. **Split the Data into Training and Test Sets Initially**:
    - Start by splitting the data into a **training set** and a **test set** at the beginning. For example, you could use an 80-20 split: 80% of the data for training (e.g., 4457 samples) and 20% for testing (e.g., 1115 samples).
    - The test set will remain unused until the very end, so you can fairly evaluate the final model's generalization ability.
    -
2. **Use Stratified Cross-Validation on the Training Set**:
    - Perform stratified 10-fold cross-validation **only on the training set** (the 80%) to evaluate different models (e.g., Logistic Regression, SVM, Random Forest).
    - Stratified cross-validation will ensure each fold maintains the proportion of spam and ham samples, which is especially helpful for imbalanced datasets like this.
    - Select the model with the best performance across these folds. Suppose, as you said, Random Forest emerges as the best.
    -
3. **Train the Final Model on the Entire Training Set**:
    - Once you've identified Random Forest as the best model, retrain it on the entire 80% training set.
    - This allows the model to learn from all available training data, which can improve its performance further.
    -
4. **Evaluate the Final Model on the Test Set**:
    - Finally, evaluate the fully trained Random Forest model on the held-out 20% test set.
    - This will give you an unbiased estimate of how well the model performs on unseen data.