Homework - 2

Question 4.1 (a)

Typically, there are mainly three types of data splitting methods.

1. **Random Splitting:** In this method, we split the data randomly thereby protecting the data from being biased. However, we cannot use this method when the response is not evenly distributed across the outcome
2. **Stratified Random Splitting:** In this method, we randomly select samples from each class thereby making sure that the frequency distribution of the outcome is approximately equal within the training and test sets
3. **Non-Random Splitting:** This method is used when there is an important temporal aspect to the data.

For the music genre dataset, we observe that there are six different types of music genres of data, each of which is not evenly distributed and have high dependency on each other. This eliminates the possibility of using random splitting. We can split the data for each class, in this case genres and then randomly select samples from them. This will ensure random selection of even data. Hence, we use **Stratified Random Splitting** for this dataset.

We can also use **K-fold Cross Validation** as a resampling technique as it will be less computationally challenging compared to bootstrapping in which a lot of computational power is required. Also, we have enough amount of data thereby eliminating the need for Bootstrapping which reuses the same rows again.

Question 4.1 (b)

(b) Using tools described in this chapter, provide code for implementing your approach(es).

To implement **K fold Cross Validation,** we can use the below R code:

```
set.seed(1)
cvSplits <- createFolds(trainClasses, k = 10, returnTrain = TRUE)
```

createFolds in the caret package can be used for data partitioning into k-folds. The above code could be used to create 10 folds.

Question 4.3 (a)

Using the “one-standard error” method, what number of PLS components provides the most parsimonious model?

	Resampled R^2	
Components	Mean	Std. Error
1	0.444	0.0272
2	0.500	0.0298
3	0.533	0.0302
4	0.545	0.0308
5	0.542	0.0322
6	0.537	0.0327
7	0.534	0.0333
8	0.534	0.0330
9	0.520	0.0326
10	0.507	0.0324

On checking the given table, we can see that having **4 components** has the maximum mean. However, we can also view that there is not a huge change in the mean if we take into consideration **3 components** instead. Hence, we can say that even though having 4 predictors produces the best parsimonious result but having 3 predictors can produce close to parsimonious result in an easy way.

Question 4.3 (b)

Compute the tolerance values for this example. If a 10 % loss in R2 is acceptable, then what is the optimal number of PLS components?

We can calculate tolerance value using the below formula:

Tolerance Value = [Mean – (Max Mean)]/ Max Mean

In the above question we are given that there is a 10% loss in R2 which is equal to 0.1

```
> df
    Mean  Error    tolerance
1  0.444 0.0272 -0.185321101
2  0.500 0.0298 -0.082568807
3  0.533 0.0302 -0.022018349
4  0.545 0.0308  0.000000000
5  0.542 0.0322 -0.005504587
6  0.537 0.0327 -0.014678899
7  0.534 0.0333 -0.020183486
8  0.534 0.0330 -0.020183486
9  0.520 0.0326 -0.045871560
10 0.507 0.0324 -0.069724771
>
```

We see that the lowest setting that does not exceed a 10% tolerance is a **2-component model**.

Question 4.3 (c)

Several other models (discussed in Part II) with varying degrees of complexity were trained and tuned and the results are presented in Fig. 4.13. If the goal is to select the model that optimizes R2, then which model(s) would you choose, and why?

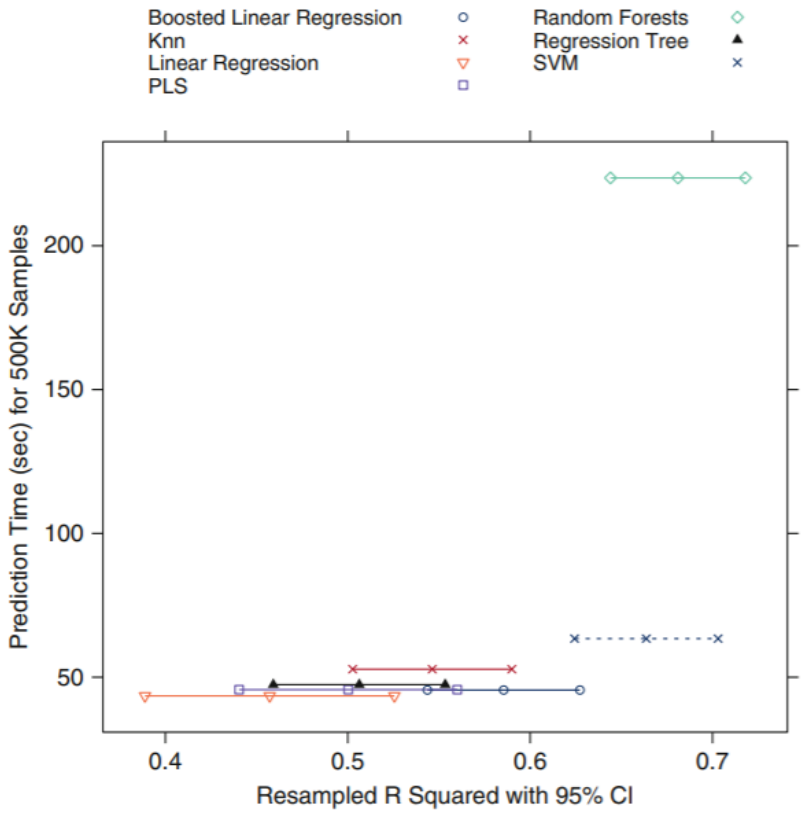

Fig. 4.13: A plot of the estimated model performance against the time to predict 500,000 new samples using the chemical manufacturing data

By taking into consideration only R^2 as a factor to choose the model we observe that **Random Forest** has the largest R^2. However**, Support Vector Machine** has almost same results and the confidence intervals for R^2 have some overlap. Hence, we can choose both models for optimization.

Question 4.3 (d)

Prediction time, as well as model complexity (Sect. 4.8) are other factors to consider when selecting the optimal model(s). Given each model's prediction time, model complexity, and R2 estimates, which model(s) would you choose, and why?

Taking into consideration factors such as prediction time, model complexity and R^2, by looking at the graph we can easily conclude that **SVM is the best model**. Even though Random Forest produces the best accuracy however, it is very time consuming when taken other factors into consideration. SVM produces similar accuracy with less prediction time thereby making it faster and less complex.