Homework Assignment for Chapter 4

4.1. Consider the music genre data.

a) What data splitting methods would you use for these data?

The music genre dataset has 12,495 samples (n) and 191 predictors depicting continuous characteristics. The data contains one independent response variable containing 6 categories of music genres. With the sample size (n) being large, then we consider splitting the dataset into training and testing sets. With the target variable being unbalanced, we consider **stratified random sampling** in splitting the data into the training and testing data sets. I would consider splitting the data 75% for the training and the remaining 25% for the testing. In tuning the parameters, a single k-fold cross validation would be considered with k=10.

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

To implement this approach, I would employ the use of the **createDataPartition** function found in Caret package. This splits the dataset into k-fold, with each fold maintaining the distribution of the classes from the original dataset. The code used to create the 10-fold would be:

Using the training/test split's function:

```
>library(caret)

>trainR<- createDataPartition(Classes, p=0.75, list=FALSE)
>trainClass<-classes[trainR]

> set. seed(1)

> Split_data<- createDataPartition(trainClass, k = 10, returnTrain = TRUE)
```

4.3. Partial least squares

a) Using the one standard error method, what number of PLS components provides the most parsimonious model?

The one-standard error method would choose the best and simplest model with accuracy not less than (0.545-0.0308)=0.5142(Lower Boundary)

From the table and given the lower boundary, it points out 3PLS which is 0.533.This is the simplest model. Therefore, best solution is component 3PLS, which is the most parsimonious model.

b) Compute the tolerance values for this example.

Tolerance is given by:

(X-O)/O

Where:

X = Performance value

O = Optimal value

	Resampled	R^2	
Components	Mean values	Standard Error	Tolerance
1	0.4440	0.0272	-0.1853
2	0.5000	0.0298	-0.0826
3	0.5330	0.0302	-0.0220
4	**0.5450**	0.0308	**0.0000**
5	0.5420	0.0322	-0.0055
6	0.5370	0.0327	-0.0147
7	0.5340	0.0333	-0.0202
8	0.5340	0.0330	-0.0202
9	0.5200	0.0326	-0.0459
10	0.5070	0.0326	-0.0697
	Optimal value=0.545		

With an acceptance of 10% loss, then the best optimal value of PLS components is a 2PLS.

c) Which model would you choose?

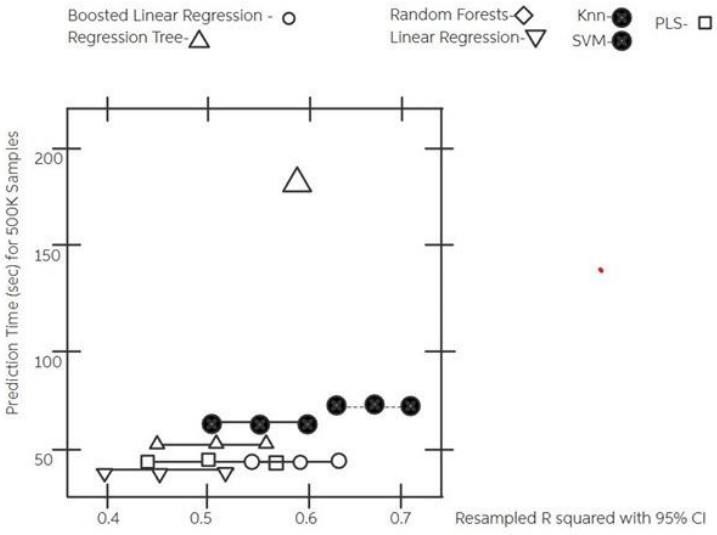

Figure 1.0

Looking at Figure 1.0, random forest is the best model with R^2. Support Vector Machine(SVM) has almost similar outcome with some overlap. Boosted Linear regression is the next possible model, but its value is worse compared to SVM. Therefore, the best models in relation to R^2 are Random Forest and Support Vector Machine.

d)Prediction time

In terms of R^2 only, SVM and random forest would be the best models to choose. However, looking at each model's R^2, model complexity and prediction time, SVM would be the best model since it is relatively fast and close to the best value of R^2. If there is the need for the predictive function to be recorded, regression tree and PLS would be considered. However, they give a low value of R^2. The best model would depend on the implementation.