# Blood Brain

Amin Fesharaki

5/20/2022

## Quesiton 2) The caret package contains a QSAR data set from Mente and Lombardo (2005). Here, the ability of a chemical to permeate the blood-brain barrier was experimentally determined for 208 compounds. 134 descriptors were measured for each compound.

## A) Load the data

## B) Do any individual predictors have degenerate distributions?

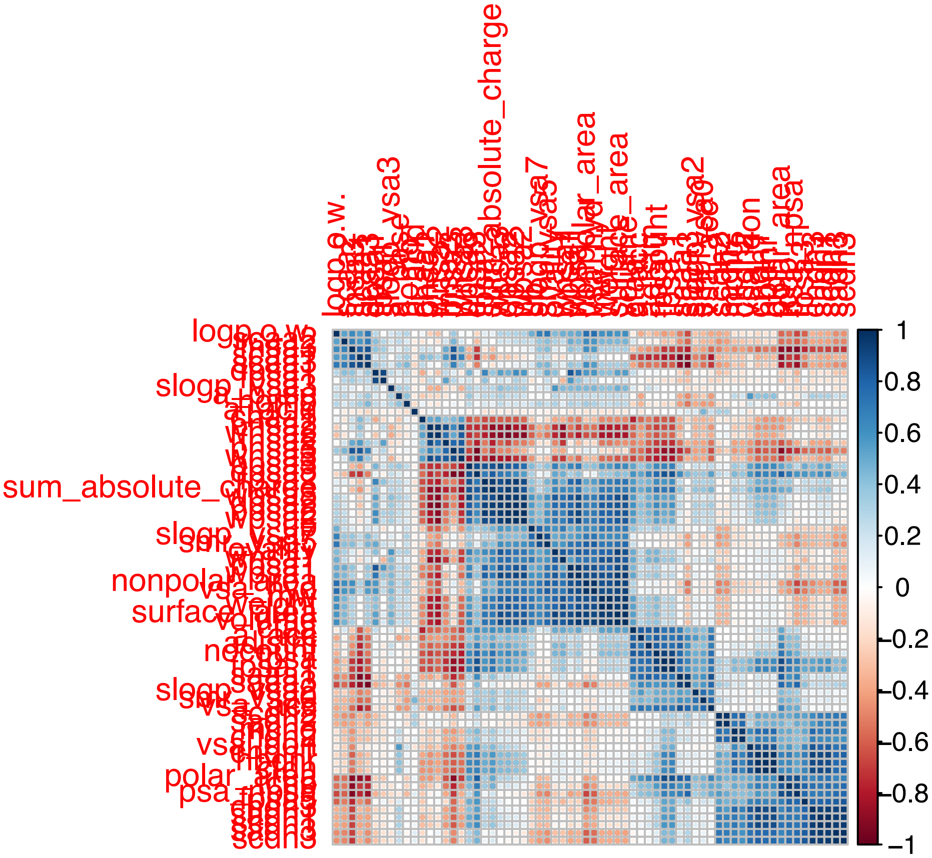
```
#Compute Near Zero Variance for all predictors
NZV <- nearZeroVar(bbbDescr, names = TRUE, saveMetrics = TRUE)
NZV_Predictors <- NZV %>% filter_all(any_vars(. %in% c('TRUE')))
NZV_Predictors
```

```
##              freqRatio percentUnique zeroVar  nzv
## negative     207.00000     0.9615385   FALSE TRUE
## peoe_vsa.2.1  25.57143     5.7692308   FALSE TRUE
## peoe_vsa.3.1  21.00000     7.2115385   FALSE TRUE
## a_acid        33.50000     1.4423077   FALSE TRUE
## vsa_acid      33.50000     1.4423077   FALSE TRUE
## frac.anion7.  47.75000     5.7692308   FALSE TRUE
## alert        103.00000     0.9615385   FALSE TRUE
```

- The table above lists the predictors with degenerate distributions (Near Zero Variance).

## C) Generally speaking, are there strong relationships between the predictor data? If so, how could correlations in the predictor set be reduced? Does this have a dramatic effect on the number of predictors available for modeling?

```
corr <- cor(bbbDescr) #Find Correlations
#Data_corr <- corrplot(corr, order = "hclust") #Correlation Plot of entire data set
StrongCorr <- findCorrelation(corr, cutoff = .75) #Indicate High Correlations with a cutoff at .75
StrongPredictors <- bbbDescr[, StrongCorr] #Extract only highly correlated predictors
Length <- length(StrongPredictors) #Return how many strong predictors (66 Highly Correlated Predictors)
StrongPredictor_Cor <- cor(StrongPredictors)
High <- corrplot(StrongPredictor_Cor, order = "hclust") #Corr plot of highly correlated predictors
```



- A correlation test with a strong correlation cutoff at .75 identified 66 predictors with high correlation. Removing 66 out of 134 predictors (49%) will have a dramatic effect on the modeling process since roughly half of the predictors will be available. A few other ways to reduce the number of correlated variables is to combine predictors together (if possible), remove linear combination predictors, and removing redundant variables. These methods will not dramatically affect the model. Another way to reduce predictors is to perform PCA to create new variables that remove the collinearity between them. However. the downside to PCA will result in variables that are difficult to interpret.