1 Cross Validation

Cross validation is a model validation technique, where the aim is to measure how accurate a predictive model performs on an unknown data set.

To start with, partition the *known* data set into two subsets, in practice a 75%/25% division is used. The 75% subset is used for training a model, the 25% subset for validation. It is important that the two subsets follow the same distribution, e.g. that the 0/1-proportion of the *return* variabel in each subset equals the 0/1 proportion of the whole set.

```
library(caret)
split.idx   <- createDataPartition(y = known$return,
        p = 0.75, list = FALSE)
split.train <- known[split.idx,]    #75% training set
split.test  <- known[-split.idx,]   #25% test set
```

In a next step, partition the 75% training subset randomly into k-groups of same size (*k-fold cross validation*). Then the model gets trained on $k-1$ groups and validated on the remaining one. This procedure is repeated k-times, such that each group serves once as validation set.

To further improve predictive accuracy, this process is repeated n-times, that is the training set gets partitioned again into k-groups (that differ in their composition from the previous group compositions), and again each of the k-groups serve once as validation set and the remaining groups as training sets.

Then apply the trained model to the 25% share test set and compare the predicted values to the true values (out of sample validation).