Voraussage Modelen von Lift Gebrauch

Preliminaries

You need to have installed the following packages, so that they can be loaded in the following chunk.

```
library(tidyverse)
library(ggplot2)
library(randomForest)
library(missForest)
```

Set your working directory to the directory where this .Rmd file resides along with the relevant .csv file, 'eintritt2.csv'. This file has the data for the exercise and was created in a previous wrangling script.

```
setwd("/Users/bgppermp/docs/school_of_data/hackathons/tourism_hackadays/Arosa_Lenzerheide/Daten_Arosa/p
```

Cross-validation of a Random Forest model

Read in the input dataset that was previously assembled.

```
set.seed(12345)
data3 <- read_csv("./eintritt2.csv")
data3 <- as.data.frame(unclass(data3))

data3 <- data3 %>%
  select(., Wetter, auslastung, Arosa, Lenzerheide, anzahlschuler, wochenend, monat, wochedesjahres, da
```

Impute missing values

The data set has 35 missing values distributed throughout the matrix. Let's use a Random Forest to impute those values, so that we don't end up throwing out valuable rows of data.

```
data3 <- missForest(data3)$ximp
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
##   missForest iteration 6 in progress...done!
##   missForest iteration 7 in progress...done!
```

```
n <- floor(0.8*dim(data3)[1])
```

Now the number of missing values is 0, so all the missing values were successfully imputed.

Now we can do a 10-fold cross-validation of the predictive performance of a random forest model.

```
form <- formula(auslastung ~ Wetter + wochenend + wochedesjahres + dayofseason
                + schneefall + schulferienindex + percentss +tag2schnee)
cross.val.rf <- function(data,form,ntrees=1000) {
  best <- NA
  Rsq <- numeric()  # a vector to store Rsq values
```

```
  #create a vector to specify fold membership of observations in dataset
  folds <- rep_len(1:10,dim(data3)[1])
  folds <- base::sample(folds,length(folds),replace=FALSE) #randomize the order of
                                                           #the folds vector
  for (i in 1:10){
    test <- data[which(folds==i),] #pull out the test data for the fold
    train <- data[which(folds!=i),]  # pull out the training data (i.e. not test)
    mod <- randomForest(form,data=train,ntree=ntrees)
    preds <- predict(mod,newdata = test)
    lmod <- lm(test$auslastung ~ preds)
    Rsq <- c(summary(lmod)$r.squared,Rsq)
  }
  return(mean(Rsq))
}

# fit a model using all the data. This should be the best model
best.rf <- function(inputdata,form){
  mod <- randomForest(form,data=inputdata,ntree=1000)
}

# determine the R-square via 10-fold cross-validation
cross.val.Rsq <- cross.val.rf(data = data3,form = form)
```

The 10-fold cross-validation R^2 of Random Forest models is 0.87. This is to say that the Random Forest model accounts for 87% of the day-to-day variation in first-time lift ridership at the ski areas.