Homework Assignment 2

Data Science and Machine Learning 1 - CEU 2018

Due date: 2018-02-05 8:00 a.m.

General information

The due date of this homework assignment is 2018-02-05 8:00 a.m. You are required to submit two files to Moodle: an `.Rmd` file and the rendered `.html` file with your solutions. Please also include code to the `.html` as well (use the default settings).

Please give short (2-3 sentences) interpretations, explanations to your answers, not only the program code and outputs.

Grading: 13 points can be earned in total. However, 10 points count as 100%. You can earn more than 100%, that is, giving a perfect answer to all questions means 130%.

```
library(data.table)
library(caret)
```

1. Predicting mental health problems in the tech sector (8 points)

In this task you are going to predict mental illness for workers in the tech sector. The data comes from Kaggle. The variable to predict is `treatment`.

```
data <- fread("../../data/mental-health-in-tech/survey_cleaned.csv")

data <- data[ ,c("comments", "state","work_interfere") := NULL]
data[, age := as.numeric(age)]
data[ , treatment := factor(treatment, levels = c("Yes", "No"))]
```

a) Explore some predictors that can be used to predict `treatment`.

b) Partition your data to 70% training and 30% test samples.

c) Build models with `glmnet` and `rpart` that predict the binary outcome of `treatment` (you don't have to use all variables if you don't want to - experiment! Just use the same variables for both model families). Use cross-validation on the training set and use AUC as a selection measure (use `metric = "ROC"` in `train` and also don't forget to use `classProbs = TRUE, summaryFunction = twoClassSummary` in `trainControl`). Make sure to set the same seed before each call to `train`.

d) Compare models based on their predictive performance based on the cross-validation information (you can just use the mean AUC to select the best model).

e) Evaluate the best model on the test set: draw an ROC curve and calculate and interpret the AUC.

f) If you have to choose a probability threshold to predict the outcome, what would you choose? At this threshold, how large are the true positive rate and the false positive rate? How many false positives and false negatives there are in the test sample?

2. Transformed scores (5 points)

Take the medical appointment no-show dataset we used in class and apply all the cleaning steps we did, then create a training and a test set. Estimate a predictive model of your choice for `no_show` as a target variable. Get predicted scores (probabilities). Then calculate two transformations of the scores: take the square root

and the square of the probabilities. These are valid scores as well, they are alse between 0 and 1 so they can be used for classification.

```
data <- fread("../../data/medical-appointments-no-show/no-show-data.csv")

# [... apply the cleaning steps we did in class ...]
# [... create train and test sets ... ]

model <- train(...)

prediction <- predict.train(model, newdata = data_test, type = "prob")
prediction_sqrt <- sqrt(prediction)
prediction_sq <- prediction^2
```

a) Draw ROC curves for all three scores and calculate the AUC. How do they compare? Is it surprising in light of the interpretation of the AUC?
b) What is the key, common property of both the square root and the square functions that leads to this finding?
c) Draw a calibration plot for all three scores separately:

- group people into bins based on predicted scores
- display on a scatterplot the mean of the predicted scores versus the actual share of people surviving

How do they compare? Which score(s) can be regarded as well-calibrated probabilites?