# NPFL054 Introduction to Machine Learning

## Charles University, November 2019

## Final Homework Exercise

---

# A) GENERAL DESCRIPTION OF THE ML TASK AND THE PROVIDED DATA SETS

## Introduction

The main objective of this homework is to solve a real-world business problem by using machine learning approaches. The goal is to build models that would be able to predict if customers are interested in purchasing caravan insurance policy or not, based on available data about customers.

## Dataset

The data source used in this homework was provided by the Dutch data mining company Sentient Machine Research. This data set contains information about its customers and the results of the marketing campaign that tells if customers were interested in a caravan insurance policy or not. The `Caravan` data set you will work with is available in R package ISLR, attached to textbook *Introduction to Statistical Learning in R* by James et al (2013).

Development data set D contains 5,822 real customer examples. Each example consists of 86 attributes, containing sociodemographic data (attributes 1–43) and product ownership (attributes 44–85). The target attribute 86 (`Purchase`) indicates whether the customer purchased a caravan insurance policy.

You are also provided with a 'blind' test data set T containing 1,000 examples without true classifications. Examples in T have been randomly selected from the same population as development data set D. Using the development examples in D you should do all your experiments and tune and compare your machine learning models. Finally you should develop and choose your 'best' model and use it for prediction on the test set T. Then you will submit your predictions as a vector of binary values and it will be evaluated. Test set T is posted at

http://ufal.mff.cuni.cz/courses/npfl054/homeworks.

Please note that this exercise is *data closed*, which means that you are allowed to work ONLY with the provided data sets.

## Features

The customers are described by 86 attributes. Sheet `caravan.attributes.pdf` posted at
`http://ufal.mff.cuni.cz/courses/npfl054/homeworks`
presents all attributes, their names, descriptions and values. Integer values refer to several defined levels, e.g. L1 levels are the values of attribute `MGEMLEEF`.

## Goal

For a while, become an insurance agent with a limited budget that allows visits to only 100 potential customers. No doubt you wish to sign as many contracts for caravan insurance policy as possible. You have records about a set of 1,000 potential customers and you want to select 100 most promising ones. To build a model for prediction you can use a known set of other 5,822 records with the information if caravan insurance policy was purchased or not.

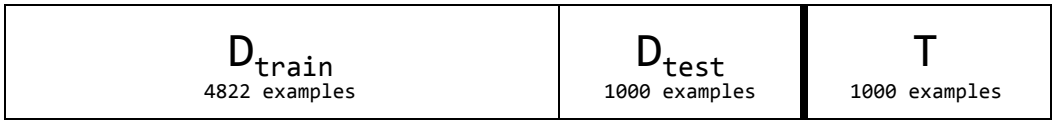
# B) EXACT SPECIFICATION OF YOUR TASKS

Load the ISLR package and look at Caravan data set.

```
> library(ISLR)
> str(Caravan)
```

## Data division

- Split the development data set D into two parts:
  - development test set $D_{test}$ – randomly selected 1,000 examples
  - development working data set $D_{train}$ – the remaining examples
- $D_{train}$ should be used for learning during the development cycle. You will tune the learning parameters using cross-validation.
- $D_{test}$ should be used only when your models are tuned to estimate generalization error and to select your best model. In fact, using the development test set $D_{test}$ you simulate the final situation when your model is evaluated using the blind test set T.
- Finally, after all development procedures, use whole development data set D for training. This will be your final model that you will use for the final prediction on blind test data T.



## Technical hints

- To learn your predictors use the following R packages: `rpart` for Decision Trees, `randomForest` for ensemble learning, and `glmnet` for (regularized) logistic regression.
- Variable importance values can be computed by `importance()` function implemented in `randomForest` package.
- To draw ROC curves and to compute AUC values we recommend using `ROCR` package and function `performance()`.
- Of course, you can use any other libraries freely available in R.

## Task 1 – Data analysis

First, check the distribution of the target attribute. What would be your precision if you select 100 examples by chance?

1a) Focus on the customer type `MOSHOOFD`: create a table with the number of customers that belong to each of 10 L2 groups and the percentage of customers that purchased a caravan insurance policy in each group. Comment the figures in the table. Then do the same for the customer subtype `MOSTYPE` (41 subgroups defined in L1).

1b) Analyze the relationship between features `MOSHOOFD` and `MOSTYPE`.

## Task 2 – Model fitting, optimization, and selection

**Important general remarks on evaluation**

- Everytime when you run a cross-validation process, you should *randomly* divide your data into a given number of folds. Since the positive and negative examples are highly *unbalanced*, you should make the division *carefully*. You should always keep the *identical number of positive examples in all your folds.*
- You should keep in mind the purpose of this classification task. Your models for binary classification should prefer high precision. Hence to optimize your model you will try to maximize a particular area under the ROC curve rather than to simply minimize error rate. Since precision naturally decreases with increasing FPR, you will measure and optimize the AUC just up to FPR ≤ 20%, hereafter denoted by $AUC_{0.2}$.

  *Technical hint*: Using function `performance()` for computing AUC you can apply and set parameter `fpr.stop=0.2`.
- Think about the particular threshold FPR ≤ 20%, how it relates to precision. If FPR was more than 20%, what could be precision then?
- Whenever you compute confidence intervals (CI) for mean in this exercise, use CI based on t-test and set the significance level $\alpha$ = 5%.

**You should develop, tune, and compare the following classification models**

2a) Decision Trees

2b) Random Forests

2c) Regularized Logistic Regression

- To evaluate each model with particular learning parameters, do 10-fold cross-validation using $D_{train}$ and measure $AUC_{0.2}$. You should optimize the following parameters:

| **Method** | **Learning parameters** |
|---|---|
| Decision Tree | complexity parameter cp |
| Random Forests | number of trees (ntree), feature sample size (mtry) |
| Regularized Logistic Regression | regularization parameter lambda, elasticity parameter alpha |

  For each learning parameter, describe how and why you selected the optimal value. Always visualize how the cross-validation result depends on the parameter using a table or plot.

  You should always compute and report the mean of $AUC_{0.2}$, its standard deviation, and confidence interval for the mean. Results should be arranged in a nice table. Confidence intervals should be used to check if the differences are statistically significant.

2d) When learning parameters are optimized, use the development test set $D_{test}$ and evaluate your tuned models using ROC curves. Compare all models and make conclusions.

2e) Finally, when you compare your models and choose the best one, you should set an optimal cut-off threshold to optimize the required precision for selecting 100 potential customers.

## Task 3 – Model interpretation and feature selection

Observing your developed predictors, discuss the feature importance. Use the Lasso model to find a reduced feature subset and compare it

- with the features used in your Decision Tree model
- with variable importance measure produced by your Random Forest model

### Task 4 – Final prediction on the blind test set

Your final model should be optimized for your main goal: to select 100 most promising potential customers out of given 1,000 blind examples in the test set T. Then you will submit the selected 100 examples to the contact teacher, who will compute your precision. The better precision value, the better predictor!

- Important technical remark
  The classification vector should be submitted in a plain text file named `T.prediction`. The number of lines in the submitted file should be just 1,000 (= the number of examples in the test data set T). There should be just one number ( 0/1 ) on each line.

## C) FINAL REMARKS

The hard deadline for your submission is **January 6, 2020**. You should submit a report with your solution by email to `holub@ufal.mff.cuni.cz` . Later submission will be penalized.

### Important directions

Make just one zip file named surname_name_hw3.zip . The submitted zip file should contain the following obligatory items:

- Text document (pdf file) with all answers, tables, plots, etc.
  - Be as clear as possible and explicitly answer all explicit questions required in this specification.
  - Do not change the order of the questions/subtasks in this specification.
  - Each Part should start at a new page.
- All source files you used to prepare the pdf document (.odt, .docx, .ods, .xlsx, .tex, etc.)
- All R codes that you wrote and used
- Any other illustrative attachments/appendices of reasonable size, if you want/need
- The classification vector in the strictly required format (see Part … )

## Scoring

In total you can get 50 points at maximum.

| Task | Points |
|---|---|
| 1a + 1b | 4 + 5 |
| 2a + 2b + 2c + 2d + 2e | 5 + 5 + 5 + 6 + 4 |
| 3 | 6 |
| 4 | 5 |

Moreover, we add up to 5 points are for the overall assessment of your report – its good structure, clarity, nice (typo)graphics, and language.