## Project Work (30%)

Project work is a chance for you to put to practice the knowledge that you have accumulated throughout the course.

Every group has received a dataset. Use data dictionary that you have got together with the allocated dataset to get familiar with the data and understand your classification task.

Familiarize yourself with the data, conduct exploratory analysis and necessary pre-processing. Use the preprocessed data, built models and communicate the results in the form of a blog-post.

Below, you can find the minimal programming requirements (what you should do in R), as well as the description of the files which you should submit and present your work. Feel free to use all the available materials (solutions to previous HWs and Rmd-files from tutorials).

## What you should do in R

## Task

1. Conduct Exploratory Data Analysis (EDA), inspect the data set and preprocess the data:

Xtreat missing values, *if applicable*

Xhandle sparse classes of categorical predictors, *if applicable*

Xtake care of outliers, *if applicable*

Xtreat skewed distributions, *if applicable*

* + explore the relationships between predictors and the target

Xcreate new features

Write a short summary (one-two paragraphs) describing the problems that you spotted and methods applied to treat them.

1. Discuss your insights from EDA, as well as what problems need treatment, with a group partner. Preprocess the data. ??Double preprocession
2. Build appropriate models, justify your choice, and make predictions:

Xvary methods

Xlogistic regression

Xlogistic regression with regularization: lasso, ridge

* + 1. decision trees)
  + try different sets of predictors

1. Evaluate the quality of predictions using at least the following metrics:

XBrier Score (RMSE)

* + AUC

## Files to submit

Since data science is not only about extracting new knowledge from the data, but also about communicating the results of the analysis, we ask you to

1. report the results of modeling in form of a blogpost,
2. prepare a short presentation (5 minutes).

## Blogpost

Compared to standard HW submissions, this report (which must also be written in R-Markdown and kneeted to Word- or PDF-file) should be pretty (graph labels, nice colors, alignment of objects, etc.) and informative. Avoid unnecessary visualizations and tables. Make sure that the output is nested in the text and not vice versa (the case of a typical HW submission).

* Start with a brief description of your dataset.
* Then summarize the most valuable insights from EDA and the problems you treated during the data preprocessing. There is no need to duplicate the last graded homework.
* Explain what methods you applied to train models.
* Present the best – in terms of prediction quality - model (how it was trained: method, features, mention meta-parameters if applicanble).
* Conclude with your personal estimate of how well you managed to predict the response. What challenges you have faced.

**Submission Deadline: Friday, 02.07.2021 at 18:00.**

## Presentation

Even though the presentation is very short, make it informative and exciting for you to present and for your group-mates as well as the instructors to listen to. We don’t provide a template on purpose, so that you can express your presentation style or work out one. Just keep in mind that it should be a business/academic presentation. For tips on how to make an excellent presentation, we encourage you to read the blog post [“Presentation Tips for Data Professionals”](https://data36.com/presentation-tips-for-data-professionals/) (click on the title to open).

During the presentation which will take place during the last class (the 14th of July), you will have to cover the following items very briefly *(note that most of them coincide with the recommended structure of a blogpost)*:

* + your prediction task
  + most exciting discoveries from the EDA (1-2 nicely looking and informative graphs, no more)
  + summary of the data “problems” that you have taken care of (missings, outliers, skewed distributions, etc., from the data preprocessing step)
  + if applicable, the features (variables) that you have additionally created using the original data
  + models you have estimated: how many in total, how they are different
  + show that you have/have not beaten the benchmark of naive predictions
  + models’ predictive quality (Brier Score (RMSE), AUC - is there an unconditional model-leader?)

**Presentation Day: Tuesday, 06.07. 2021 during the class via Zoom (the exact time will follow).**