* check if there are features we need to remove: bad distribution, too much missing data,..
* missing value ?
* change data representation: one hot, change numeric to interval?,... (data wrangling, binary to category?)
* check existing code in kaggle
* identify outliers
* make shuffle of the dataset
* separate train/validation/test with good distribution => 80 % train+validation (k-fold cross validation) and 20%test => 20 000 test and 70 000 train/validation
* build simple model
* analyze result of the model: find good way to measure results and maybe idea how to improve it