* check if there are features we need to remove: bad distribution, too much missing data, not important based on shap
* missing value ?
* identify outliers
* 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
* analyze result of the model: find good way to measure results and maybe idea how to improve it