I begin by importing the data to pandas data frame and did preliminary data analysis to find out how does it look. Applying the pandas discretion function showed the distribution of each continues variables. The data set was made up of variables with numerical values as well as categorical. Looked at distribution of the continues numerical variable by creating a histogram, they were normal distribution. I looked at number of data point for each outcome. We want to have around equal, but this dataset wasn’t equal.

After the exploration I made “corrections” to the data. Since variables were normally distributed, I didn’t do anything there. Unequal number of data points could lead to bias so I did over and under-sampling to try both, under-sampling creates loss of information and over-sampling leads to overfitting. I did the over and under-sampling on the train dataset, so as to not have test dataset in the train dataset. The modes only except numerical inputs, so I turned non-numerical (categorical) into numerical by means of creating pandas dummy variable function.

Now the data is ready to go into the model. The data was small enough and not many parameters to vary, here I decided to do a grid search with sklearn gridsearchcv by varying few parameters to optimize the model. I applied this grid to both over and under-sampled training dataset and trained two models. I tested the models with one test dataset created at the beginning. I looed the cross validation with 5 kfolds, they were perfect, the result didn’t vary with the kfold and compare to the score of training dataset. I would have tried different models here but did not seem like this data set needed it.