Part 1 - Surgery duration prediction

In order to find the most relevant features, I first looked at the ID of the doctors and anesthetists. Almost all doctors and anesthetists appear only once in the data, there are about fifty who appear twice and one anesthetist appeared three times, so it seems that it is not possible to learn about the abilities of a specific doctor compared to another doctor.

I thought that there might be a correlation between the ID and the experience of the doctor (low number - more experience) but I did not see such a correlation between the number and the length of the operation.

So both IDs are ignored.

(Code for counting how many time each doctor appears: ***df.DoctorID.value\_counts().value\_counts()*** )

To check the relevance of type of surgery and of type of anesthesia, I divided the information into groups, once according to the type of surgery (five groups), once according to the type of anesthesia (two groups) and once according to both. You can see their distribution in the ***Age – BMI – Surgery-Time*** space in the following graphs:

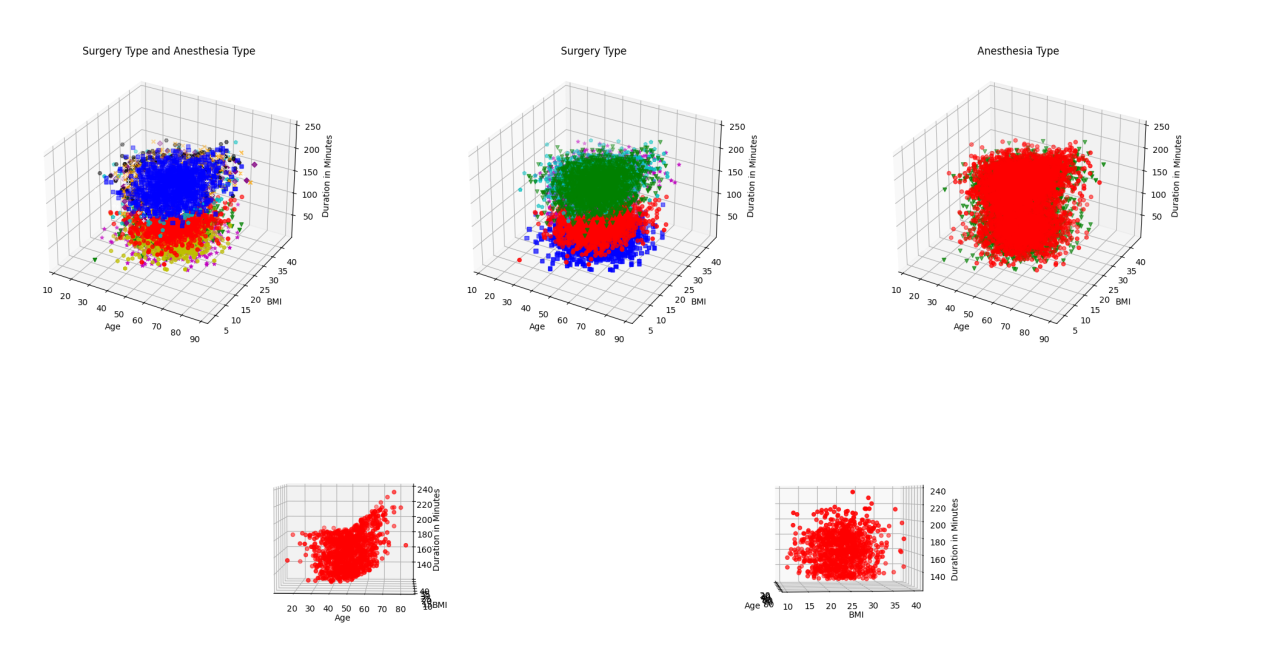


Fig 1From left to right (top) - division into ten groups (type of anesthesia and type of surgery), you can see differences between the groups. Division into five groups (type of surgery) differences can also be seen. And division into two groups of the type of anesthesia - there does not seem to be any differences.

(Bottom) - Presentation of one group with the same type of surgery and the same type of anesthesia, the two graphs are the same but rotated differently so that we can see which feature is more relevant - it seems that age has a greater meaning than BMI.

The metric I chose to evaluate the success of the regression is RMSE because it "punishes" more for a large inaccuracy, for example an error of half an hour (which causes an surgery to be postponed by half an hour) is more annoying than two delays of fifteen minutes.

It can be seen from the above figure that the type of surgery and the age of the patient are significant, in the case of the age of the patient it looks like some sort of quadratic function.

I've divided the data to training, validation and test (60-20-20) the 20% of test samples, are fixed so that we do not touch the information of the test while training, the division between training and validation is not fixed to avoid over fitting, so the following results may vary between runs. (the results are on the validation data)

I tried the following methods for estimation the surgery duration: 1. The estimation is the average of all surgeries – very naïve estimation. **RMSE = 54.47**

2. The estimation is an average of the surgery duration in the group, by group I mean all the surgeries that are of the same type and same anesthesia. **RMSE = 19.78**

3. The estimation uses k nearest neighbors on the relevant group, where the distances are age and BMI after both are normalized. I tried it at multiple k values, the best was with k=101. RMSE = **19.16**

4. Like 3 but the distance is measured only using age (and no normalization is needed). Best k was k = 61. RMSE = **18.82**

5. For each group, polynomial estimation using age. **RMSE** (degrees of polyniom = 1,2,3,4): **19.01, 18.66,18.71,18.72**

6. Gradient boosting regressor – Using OneHotEncoder I've created 7 variables for the surgery type and anesthesia type. And added the values of the age and bmi (using polynomial features which adds Age\*Age, Age\*BMI, BMI\*BMI) and run it on multiple GBRs with different learning rate, depth and number of estimators, the best result was when using 50 estimators, 0.1 learning rate and max depth of 3. **RMSE: 18.81**

7. Random forest regressor – the same data as in 6, with best result with 400 estimators and max depth of 6. **RMSE: 18.84**

It looks that the most important feature is the surgery type and after that the age, the others don't seem like adding a lot of value for the estimation.

Looking on the results, group polynomial estimation using polyfit on age only, with degree of 2 gave us the best results (although there some others that are very similar) so I run it on the test to get more realistic error estimation, and get **RMSE = 18.71**