Predicting surgical case durations for a Thorax centre

Primary Topic: DM, Secondary Topic: DPV

Course: 2022-1A 201400174 Group: 60 – Submission Date: 2022-11-13

Akram Sabik
University of Twente
s.akram@student.utwente.nl

Diana Canosa Tajes
University of Twente
d.canosatajes.student@utwente.nl

ABSTRACT

This paper tackles the problem of high rate of operating rooms working beyond regular operating time in the medical center MST in the region of Twente. Overtime causes unnecessary costs and low staff satisfaction. The goal is to find out which features are most efficient for predicting surgery duration by using data collected from patients suffering from cardiothoracic diseases from January 2013 to January 2016. As a result of our work, we show the importance of surgery data over patient data, the importance of the surgery type regarding relevant features for time prediction, and a more efficient prediction model is created and validated.

KEYWORDS

Variable Imputation, Data Science, Classification, Prediction

1 INTRODUCTION

The main task in this project is to find patterns in surgical case duration and derive prediction and classification models to efficiently plan surgery schedules for the MST medical center. This would avoid unnecessary costs, overtime for healthcare professionals and low staff satisfaction. The data set provided contains three years of surgery information (from January 2013 to January 2016). The project's research question is:

Which factors are most efficient for predicting surgery duration?

The problem is tackled using two different approaches. The first one focuses on opposing surgery and patient data to find out which one is more relevant to classify on-time surgeries. The second one's goal is to know if all surgery types share the same important variables in order to create a prediction model for surgery time duration.

Subquestion A: *Would a model using only patient data work better than a model using only surgery data?*

Subquestion B: *What features should be chosen to find a better model than the current one the hospital uses to predict surgical time?*

2 BACKGROUND

In this section the functions and methods chosen are explained.

- **feature_selection.SelectKbest (score_func: f_classif):** Select features according to the k highest scores. Part of the scikit-learn library.
  f_classif: Score function. Compute the ANOVA F-value for the provided sample.
- **over_sampling.RandomOverSampler()** Over-sample the minority class(es) by picking samples at random with replacement. Used to balance out target classes. Part of the imb-learn library.

Data encoding in an essential part of data preparation. Multiple models only accept numerical inputs. String variables must then be converted to numbers in order to use them.

- **pandas.get_dummies():** Convert categorical variable into dummy/indicator variables. Creates as many variable as there are classes (One-Hot encoding).Part of the pandas library.
- **preprocessing.LabelEncoder()** Encode target labels with value between 0 and n_classes-1. (Label encoding).Part of the scikit-learn library.
- **impute.IterativeImputer():** Multivariate imputer algorithm. Estimate each missing value taking into account the entire set of variables, in a round-robin fashion. It is inspired by the R MICE (Multivariate Imputation by Chained Equations) package. It only can be used with numerical data, so categorical variables were converted to numerical before implementing the method. Part of the scikit-learn library.

3 APPROACH

The approach we followed was firstly, to clean the data; secondly, select variables using different methods and finally validate this selection by creating prediction models.

3.1 Early decisions

In the scope of the project, early decisions were made around data cleaning before following a different roadmap for each subquestion.

The first step was to remove variables based on clinical reasoning and the research question. Variables with heavily missing data (more than 80%) were also removed.

Next, it was decided to only focus on the 4 most performed surgeries in the hospital: *CAGB* , *CABG + Pacemakerdraad tijdelijk* ,, *CAGB + AVR* (see fig.1).

Finally, the star schema was created (see fig.2) and was uploaded to the database.

3.2 Subquestion A

The idea behind this subquestion is to focus on patient data and surgery data separately. We will focus on a binary classification problem: A surgery is considered on time if it lasts the planned time ±5 minutes. Thus, 3 supervised learning models will be trained and compared in order to get to the final answer. This subquestion was solved using the following roadmap:

(1) Importation

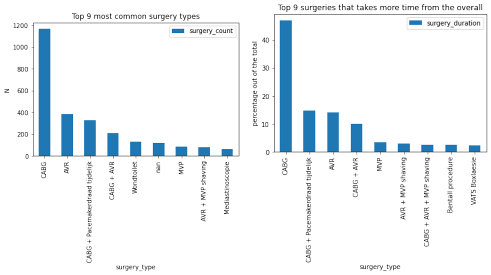

Figure 1: Top 9 Most performed surgery types

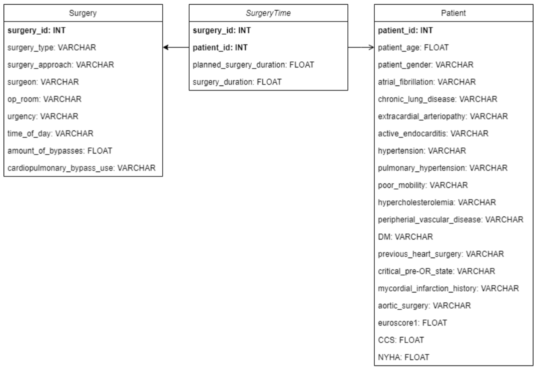
Surgery
surgery_id: INT
surgery_type: VARCHAR
surgery_approach: VARCHAR
surgeon: VARCHAR
op_room: VARCHAR
urgency: VARCHAR
time_of_day: VARCHAR
amount_of_bypasses: FLOAT
cardiopulmonary_bypass_use: VARCHAR
SurgeryTime
surgery_id: INT
patient_id: INT
planned_surgery_duration: FLOAT
surgery_duration: FLOAT
Patient
patient_id: INT
patient_age: FLOAT
patient_gender: VARCHAR
atrial_fibrillation: VARCHAR
chronic_lung_disease: VARCHAR
extracardial_arteriopathy: VARCHAR
active_endocarditis: VARCHAR
hypertension: VARCHAR
pulmonary_hypertension: VARCHAR
poor_mobility: VARCHAR
hypercholesterolemia: VARCHAR
peripherial_vascular_disease: VARCHAR
DM: VARCHAR
previous_heart_surgery: VARCHAR
critical_pre-OR_state: VARCHAR
mycordial_infarction_history: VARCHAR
aortic_surgery: VARCHAR
euroscore1: FLOAT
CCS: FLOAT
NYHA: FLOAT

Figure 2: Star schema

    1.1 Import data from database
    1.2 Merge tables
    1.3 Create output results
(2) Cleaning
    2.1 Data visualization
    2.2 Data imputation
(3) Preparation
    3.1 Data encoding
    3.2 Variable Selection
    3.3 Normalize Continuous values
    3.4 Data augmentation
(4) Classification Models
    4.1 Logistic Regression
    4.2 Decision Tree
    4.3 Neural Network

3.3 Subquestion B

What features should be chosen to find a better model than the current one the hospital uses to predict surgical time?

In this sub-question we are going to focus on predicting surgical time duration, continuous output. Instead of comparing patient variables with surgical variables as we did in SUBQUESTION A, we want to compare variables between different surgical types. The idea is to see if we can use the shared features between different surgeries to create a unique model that can predict time.

Table 1: Creation of output columns

plannedtime	*realtime*	*time_difference*	*time _difference _abs*	*overtime*
207.0	222.0	-15.0	15.0	1.0

4 EXPERIMENTS

4.1 General Setup

In this section, the general experimentation setup and solutions for each subquestion, as well as the obtained results are described.

4.1.1 SUBQUESTION A. In this section, we go over the most important steps described in the subquestion's roadmap in Section 3.

Importation: The first step is to import each table and merge it with the outcome table *suregeryTime*. The variable *overtime* takes the value 1 if the surgery is either undertime or overtime, 0 otherwise (see tab.1). Thus, we get two tables *patient_full* and *surgery_full* both including the target value.

Cleaning-Data Visualisation: The idea here is to get acquainted with the data visually. By using graphs, it might be possible to find a relationship or a trend between variables that can be useful for the next steps (see fig.3).

Cleaning-Data Imputation: Missing data is dealt with by using data imputation. Table 2 shows the different types of imputation that were used for which variable types as well as examples. One of the goal of this section was to get an accurate imputation. During the previous step, it was noticed that the variables *NYHA* and *CCS* share a relationship with the variable *patient_age* (see fig.4). It was then decided to impute the missing data using the mean value per age in order to increase the quality of the imputation.

Preparation-Data Encoding: It is important to stay mindful of how data encoding is done (see Section 2. for more details). Here, 2 types of data encoding were used (see tab.3):

- OneHot Encoding is a type of data encoding for variables with no natural order. This type of encoding allow to use numerical data without creating unwanted relationships that can lead to wrong interpretations by the models. It makes sense to use this for binary / categorical variables like *patient_gender*. The tradeoff is that this can lead to the creation of many columns when there are many classes. Sometimes it is necessary to restrict the amount of classes to avoid this problem.
- Label Encoding is for ordinal variables. These variables have an intrinsic ordering. Preserving this ordering relationship will allow the models to "understand" them more easily. For example, this type of encoding is used for the variable *amount_of_bypasses*. This variable contains an ascending relationship: The difference between 0 and 5 bypasses is greater than for 4 and 5 bypasses.

Preparation-Variable Selection: This step is especially important for building the neural network model. In this case, we select

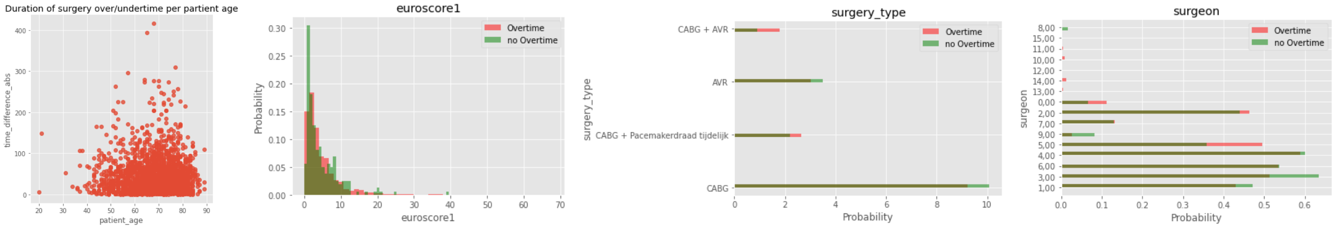

Figure 3: Early Data Visualisation: Patient and Surgery data

Table 2: Types & Examples of Data Imputing

Imputation Type	Variable type	*patient _data*	*surgery _data*
Mean per age	Special case	*NYHA* (16.1%) *CCS* (21.7%)	x
Most common value	Binary / Categorical	*patient_gender*	*surgical_approach*
Mean	Continuous	*patient_age*, *euroscore1*	x
Drop	Outcomes	*realtime*	*realtime*

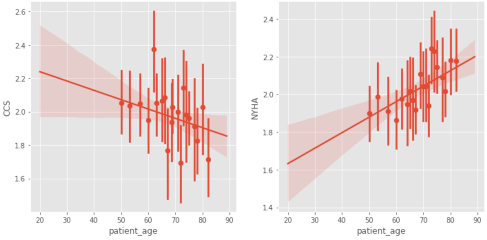

Figure 4: Early Data Visualisation: Linear Regressions

Table 3: Types & Examples of Data Encoding

Encoding Type	*patient _data*	*surgery _data*
OneHot Encoding	*atrial_fibrillation*	*cardiopulmonary_bypass_use*
Label Encoding	*pulmonary_hypertension*	*amount_of_bypasses*

the top 5 most relevant variables using the *f_classif* score function. (see Section 2. for more details)

Preparation-Balance output classes: This is an essential part for making sure that the classification models will not be biased towards the most common class. In this case, the dataset is transformed using an over-sampling method (see Section 2. for more details) to get +920% *on-time* surgeries and a 1:1 ratio in output classes.

Classification Models - Logistic Regression:
Logistic Regression is widely used for binary classification.

Classification Models - Decision Tree:
Decision Trees are a simple, but powerful form of multiple variable analysis. It performs its own feature selection using the Gini index. In this case, no depth constraint is used.

Classification Models - Neural Network:
The idea behind using a neural network (NN) is to see how a more complex model can perform using this dataset. The main challenge for using one in this context is that NNs are data greedy, and our dataset isn't that big. To solve this, 2 techniques were experimented with to obtain a better performance:

- Feature selection
- Data augmentation

4.1.2 SUBQUESTION B.

What features should be chosen to find a better model than the current one the hospital uses to predict surgical time?

CABG and *AVR* surgery types were selected among the four types we had in the star schema because they are completely different procedures. *CABG* and *AVR* procedures were included in the other two surgery types and we could have some bias in the results if we compare similar procedures.

For predicting surgical duration using data mining it is very important to gather all the significant information that the model will need. The data the hospital provided had a lot of missing data and it had to be imputed before continuing with variable selection. For reducing the errors we used multivariable imputation using IterativeImputer algorithm for several variables, as for example *CCS* or *NYHA* because these variables are cardiac scales that are dependent on other variables that were not missing in the original data.

We are looking for a continuous output, so the variables were selected to be significant for a multiple linear regression model. This step was done for *CABG* and *AVR* independently. What we did was to create a linear model with all the variables, patient and surgery, and select the ones that verify: pvalue< 0.05. We also use SelectKbest algorithm to select variables and we compared the Adjusted R squared value of both linear models, the one created using pvalues< 0.05 and the one created using SelectKbest. It was

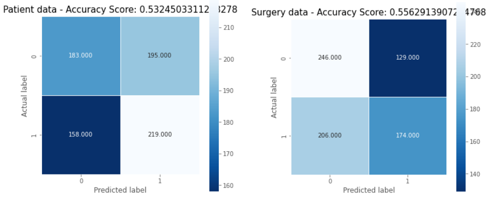

Figure 5: Confusion Matrices for logistic regression models

found that SelectKbest is a worst method to select variables for linear model as the Adjusted R squared was smaller. With this information in mind, the first method of feature selection was used for both surgery types: *CABG* and *AVR* and the results were compared to search for similarities.

To validate the variable selection, a multiple linear model was created only for *CABG* surgery. As the significant variables for each surgery were very different, we considered it not suitable to create the same model for all the surgery types. For that, the *CABG* imputed data was used again but this time it was divided into train-set (0.8) and test-set (0.2). The model needs to be tested with a different data set that was not used before to create the model. Which means that the previous variable selection cannot be used for this model as the whole data set was included to obtain the features. New variables were selected with the previous method of selecting a pvalue< 0.05.

This time, we had to deal with multicollinearity between variables when creating the linear model. Making some research we find that *Euroscore1* include the patient age [1], so we remove patient age to reduce the condition number from 1180 to 117. The Adjusted R Squared of the model was 0.362, which means that the model only predicts 36.2% of the train set.

A regression tree model was also created using the same train-set to compare different models. The comparison were performed calculating the RMSE for each model, between the predictions the model gives for the test-set surgeries and real time. For the RMSE of the current model the hospital used, the RMSE were calculated between planned time and real time of the same test-set surgeries.

4.2 Results

4.2.1 SUBQUESTION A. The 3 types of models that were developed for each dataset can be compared using selected performance metrics:

- The regression models' confusion matrices can be compared in fig.5
- The decision trees can be visualized in fig.6
- The NNs' accuracy & loss plots can be compared in 7
- Finally, all the models' performance and parameters are contained in tab.4.

4.2.2 SUBQUESTION B. Table 5 shows the results of feature selection for *CABG* and *AVR* surgery types . It can be seen that only two features are shared between both surgery types. This was the main reason for creating the regression models only for one surgery type.

The results for the different model RMSE (see Tab.6) shows a high RMSE for the regression tree, 68 minutes.

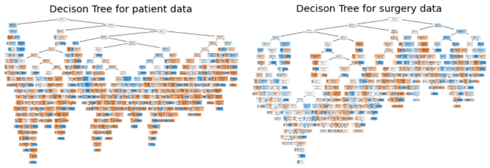

Figure 6: Decision trees models visualized

Table 4: Performance Evaluation

Model	Measure	*patient_data*	*surgery_data*
Logistic Regression	Accuracy	0.53	0.56
Decision Tree	Accuracy	0.93	0.79
	Tree Depth	21	23
	Leaves Number	379	348
Neural Network	Input	5	8
	Accuracy	0.56	0.62
	Loss	0.67	0.65
	Epochs	40	40
	Total Params	609	433

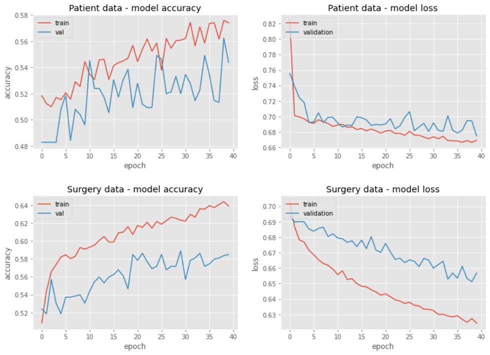

Figure 7: NN Accuracy & Loss plots

Table 5: Significant variables for multiple linear model

CABG	AVR
patient_age	*urgency*
atrial_fibrillation	*myocardial_infarction_history*
previous_heart _surgery	*CCS*
Euroscore1	*NYHA*
amount_of_bypasses	*amount_of_bypasses*
cardiopulmonary_bypass_use	*surgeon*
surgeon	
time_of_day	

It was found that the linear model created has a lower RMSE than the current system of the hospital. In Fig 8 both models can be compared. In the x axis real time is plotted and in y axis predicted time is plotted. The more diagonal the line of regression is, the

Table 6: RMSE comparison

	hospital plan	Linear Model	Regression Tree
RMSE (min)	49.4	44.7	68.1

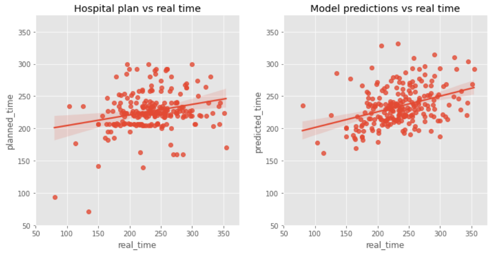

Figure 8: Comparison between hospital planned time and linear model predicted time

better the predicted time coincides with the real time, so the better the model performs. It can be seem that the multivariate linear model has a more diagonal regression line.

5 DISCUSSION

5.0.1 SUBQUESTION A. When looking at the performance of logistic regression and decision tree for each dataset, one can conclude that surgery data is more efficient to classify on-time surgeries. Out of the 3 models, the decision tree is the most accurate. However, the size of the trees obtained (especially for the patient dataset) are important as they were created using a lot of features. This makes them more difficult to visually interpret them. The neural network model using surgery data also seems more potent for the task than the one using patient data: this can be noticed when looking at the trend of the loss curve steadily going down after each epoch. However, those 2 NN models are heavily dependant on how many variables they have as input. This means that better feature selection can lead to a better performance. Data augmentation also plays a big role: In this case, +920% *on-time* surgeries need to be generated to obtain a 1:1 ratio. A bigger dataset is definitely needed for a better training of the models. According to [3], the most common causes for delays in surgeries are *Lack of proper planning, Deficiencies in team work, Communication gap , Limited availability of trained supporting staff.* This enhances the idea that surgical data plays an important role and it is advisable to collect more surgery-related data in order to improve prediction/classification capabilities.

5.0.2 SUBQUESTION B. From table 5 it can be seen that only two features are shared between both surgery types. This means that our initial idea of creating a unique model for all surgery types is not the best, and a specific linear model for each one would provide better results.

An important observation for this part is that SelectKBest is a method used for classification problems feature selection and the problem we have now is a regression problem. This coincides with our worst results in Adjusted R Squared when we used it for the variable comparison part. Another method specific for linear regression must have provided better feature selection. However, for creating the models we discard SelectKBest, so it is not going to affect the final results.

Comparing RMSE between the models can be seen that the multivariate linear model created has a root mean squared error of 44.7 minutes, whereas the hospital planning RMSE is 49.4 minutes. Figure 8 coincide with the RMSE results and show a more diagonal line for the linear model than for the scheduled time the hospital predicted. Therefore, it can be concluded that the multiple linear regression model created is going to perform better for CABG surgeries.

Even though the linear model predicts better surgical time duration for *CABG* surgeries than the one the hospital is using, it has a strong multicollinearity. When a condition number is larger than 30 the collinearity is regarded as strong [2] and our model has a condition number of 117 after removing *patient_age* variable (collinear variable with *euroscore1*). Therefore, the model can still be improved to reduce this multicollinearity that is not supossed to be present when creating a linear model.

It is also important to be aware that imputation of missing variables is introducing some errors that are going to affect the final performance of the model.

6 CONCLUSIONS

We showed thanks to 3 types of classification models that surgery data seems to be more relevant than patient data to classify on-time surgeries.

We proved that different variables are significant for each surgery type. So our initial approach of selecting the best features to create a prediction model for all surgeries was changed. Taking into account this results, a prediction model for each type of surgery seems to be the most effective way to predict surgical time.

A multivariate linear regression model for predicting *CABG* surgery time was created and proved to worked better than the one the hospital has for a test-set, so it reinforces the previous result of the suitability of creating a different model for each surgery type. However, the model still has a strong multicollinearity, so it still needs to be improved to be used by the hospital.

REFERENCES

[1] EuroSCORE Webpage.European system for cardiac operative risk evaluation.

[2] Kim JH. Multicollinearity and misleading statistical results. Korean J Anesthesiol. 2019 Dec;72(6):558-569. doi: 10.4097/kja.19087. Epub 2019 Jul 15. PMID: 31304696; PMCID: PMC6900425.

[3] Gupta B, Agrawal P, D'souza N, Soni KD. Start time delays in operating room: Different perspectives. Saudi J Anaesth. 2011 Jul;5(3):286-8. doi: 10.4103/1658-354X.84103. PMID: 21957408; PMCID: PMC3168346.