PU5558-PU5567-computer-lab-report

Anonymous

Load necessary packages

We begin by loading the nessicary packages required to carry out all comands needed in this report. Tidyverse, as we learnt in our previous module is required for our general data science, along with "ggplots" for visualisation, and corrplot for visualising correlation matrices and tinymodels which was introduced to us this term is required for machine learning - our main focus this term.

```
# install.packages("devtools")
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.5.1     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.1
## v purrr     1.0.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(tidymodels)
```

```
## -- Attaching packages -------------------------------------- tidymodels 1.3.0 --
## v broom        1.0.7     v rsample      1.3.0
## v dials        1.4.0     v tune         1.3.0
## v infer        1.0.7     v workflows    1.2.0
## v modeldata    1.4.0     v workflowsets 1.1.0
## v parsnip      1.3.1     v yardstick    1.3.2
## v recipes      1.2.1
## -- Conflicts ----------------------------------------- tidymodels_conflicts() --
## x scales::discard() masks purrr::discard()
## x dplyr::filter()   masks stats::filter()
## x recipes::fixed()  masks stringr::fixed()
## x dplyr::lag()      masks stats::lag()
## x yardstick::spec() masks readr::spec()
## x recipes::step()   masks stats::step()
```

```
library(ggplot2)
library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
# install.packages("devtools")
devtools::install_github("r-lib/conflicted")
```

```
## Skipping install of 'conflicted' from a github remote, the SHA1 (4d759ac6) has not changed since last
```

```
##   Use `force = TRUE` to force installation
```

Load chosen dataset

We were provided two seperate data sets containing "Provisional Patient Record Outcomes" to select from to answer our question for this assignment. Out of the two I have selected "Knee replacement CCG 2021" which can be found at the following address: https://digital.nhs.uk/data-and-information/publications/statistical/patient-reported-outcome-measures-proms/hip-and-knee-replacement-procedures-april-2020-to-march-2021

To load this data set into my studio to then work with I will be using the "read_csv" function.

```
#Read in data
Knee_Data <-read_csv ("~/Downloads/Knee Replacement CCG 2021.csv")
```

```
## Rows: 5422 Columns: 81
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr  (5): Provider Code, Procedure, Year, Age Band, Gender
## dbl (76): Revision Flag, Pre-Op Q Assisted, Pre-Op Q Assisted By, Pre-Op Q S...
##
## i Use `spec()` to retrieve the full column specification for this data.
## i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Dataset description

Suitable machine learning algorithm for three questions:

1. Before the operation, can we estimate the post-operative EQ5D index for a patient?

When looking to estimate post-operative EQ5D index we are looking at a regression problem as the EQ5D is a continuos value outcome, therefore require supervised machine learning to solve the issue. This task requires to use a dataset of patient reported outcomes and other relevant information to train a model to predict the EQ5D score. Therefore I would choose to utilise a random forest regressor as it has the ability to handle both catagoricol and numerical predictors and is good when handling data which may contain missing values and outliers.

When looking into which model would be best suited for the task I also investigated the use of linear regression however for this task in particular it is important the model does not assume a linear relationship between variables and allows for more flexibility when working with more complex interactions in patient data, e.g age or pre-existing conditions.

Overall, I beleive Random forest has a better balance of predictive performance with practical usability for this task making it my choice for estimating post-operative EQ5D- index.

2. Before the operation, can we predict how much pain a patient will have after the operation?

In the 'Knee replacement data set' we have the data needed to carry out this task in the column 'Knee Replacement Post-Op Q Pain' where the intensity of pain has been given a numerical value therefore making this a regression problem. Classification, a type of supervised machine learning method should be used in this task. Again linear regression could be used in this task however it would assume the linear relationship between predictors and pain score. I would be more inclined to use either Random forest regressor or Support vector regression in order to predict pain.

3. Before the operation, can we calculate how many patients have had previous surgery?

Yes, we can calculate the number of patients that had prior surgeries in our data set. Our data set contains the column "Pre-Op Q Previous Surgery" in which patients answers are coded as follows; 1 = 'Yes', 2 = 'No', and 9 = 'Missing value'. Therefore we can filter for specific '1' value to calculate the number of patients

who had previous surgies. In our data set this means that there were 537 patients in our data set who had previous surgery.

table(DZ_data$`Pre-Op Q Previous Surgery`)

Model building to answer chosen question

1. Data splitting
2. Selection and preprocessing of predictors
3. Model specification and training
4. Model evaluation

Limitations of machine learning model