Edx Capstone: Prediction model Project

Luis G. Vazquez de Lara Cisneros.

02/12/2020

Contents

1 Introduction/overview/executive summary

Covid-19 is the name of a disease caused by SARS-CoV-2, a novel type of coronavirus that has spread all over the world, presenting a severe threat to global health. The clinical presentation is very heterogeneous, ranging from an asymptomatic disease, to a life threatening condition with respiratory failure. At present, a specific therapy is lacking. Because the mortality of hospitalized patients with this disease varies from country to country, it is imperative to develop prediction models tailored to the reality of the location. In Mexico, as of 17 December 2020, the health authorities reported 1,289,298 confirmed patients, with 116,487 deaths, around 9.03% of reported cases (Secretaria de Salud 2020). It has also been noted that the mortality of hospitalized patients varies from institution to institution. *The Mexican Institute of Social Security* (IMSS, after the initials in Spanish), carries the biggest burden of public health care and the highest mortality of patients hospitalized with covid-19 (Sanchez Talanquer 2020).

Medical researchers often use prediction models as an aid to estimate the probability of risk for a specific outcome, to inform their decision making. In recent years, an initiative called The Transparent Reporting of a multivariate prediction model for Individual Prognosis Or Diagnosis (TRIPOD), made a statement consisting of a checklist with 22 items considered as essential for good reporting of these type of studies (Collins et al. 2015). This work is the final assignment of the capstone course of the *Edx Data Science Professional Certificate*, where we learned the basics of machine learning techniques. Even though the course is introductory to this topic, I believe it gives the necessary knowledge to fulfill the TRIPOD requirements concerning the statistical analysis (table 1) and presentation of results (table 2).

Table 1: TRIPOD items concerning the statistical analysis (from (Collins et al. 2015)).

Topic	Item	Checklist item
Sample Size	8	Explain how the study size was arrived at.
Missing data	9	Describe how missing data were handled (ag., complete-case analysis, single imputation, multiple imputation) with details of any imputation method.

Topic	Item	Checklist item
Statistical analysis methods	10a	Describe how predictors were handled in the analyses.
	10b	Specify type of model, all model-building procedures (including any predictor selection), and method for internal validation.
	10c	For validation, describe how the predictions were calculated.
	10d	Specify all measures used to assess model performance and, if relevant, to compare multiple models.
	10e	Describe any model updating (eg., re-calibration) arising from the validation, if done.
Development vs. validation	12	For validation, identify any differences from the development data in setting, eligibility criteria, outcome, and predictors.

Table 2: TRIPOD items concerning the presentation of results [from (Collins et al. 2015))].

Topic	Item	Checklist item
Participants	13a	Describe the flow of participants through the study, including the number of participants with and without the outcome and, if applicable, a summary of the follow-up time. A diagram may be helpful.
	13b	Describe the characteristics of the participants (basic demographics, clinical features, available predictors), including the number of participants with missing data for predictors and outcome.
	13c	For validation, show a comparison with the development data of the distribution of important variables (demographics, predictors and outcome).
Model development	14a	Specify the number of participants and outcome events in each analysis.
	14b	If done, report the unadjusted association between each candidate predictor and outcome.
Model specification	15a	Present the full prediction model to allow predictions for individuals (ie, all regression coefficients, and model intercept or baseline survival at a given time point).
	15b	Explain how to the use the prediction model.
Model performance	16	Report performance measures (with CIs) for the prediction model.

I will use a database of hospitalized patients from a Mexican hospital and try to build a mortality prediction model. The database comprises information from 642 hospital records, with 51 variables. Table 3 shows the column names of the database, with a short description of their meaning and the codification or units employed. I will also use the TRIPOD items as the framework to build the statistical analysis and the presentation of results. Besides, I will try to demonstrate the skills acquired concerning data wrangling. The key steps are summarized as follows:

- Prepare the database for a suitable statistical analysis.
- Explore the data, to decide which variables will be used as predictors (*preprocessing*). In this step, I will use bivariate analysis and imputation methods.
- Split the data set into train and test sets.
- Use logistic regression, k-nearest neighbors, classification and regression trees (CART) and random forest as prediction algorithms. I will use bootstrap for cross-validation to tune-up the parameters. I will use the `caret` package to build the algorithms.
- Compare the performance of the algorithms with the *overall accuracy* in the test set.

Table 3: Description of the variables in the working database

Column name	**Description**	**Codification/units**
sexo	Gender	0 = Female, 1 = Male
ocupacion	Occupation	1 = Health care worker, 2 = Office job, 3 = Outdoor work, 4 = Work in public area,5 = Work at home, 6 = Unemployed
escolaridad	Schooling	1 = Analphabet, 2 = Primary school, 3 = Junior high school, 4 = High school, 5 = College, 6 = Postgraduate studies
nivsoc	Socioeconomic level	1 = Low, Medium-low, 2 = Medium-high, High
has	History of High blood pressure	1 = Yes, 2 = No
tabaquismo	Smoking	1 = Yes, 2 = No
dm	History of diabetes mellitus	1 = Yes, 2 = No
renal	History of renal failure	1 = Yes, 2 = No
autoinmunidad	History of autoimmune diseases	1 = Yes, 2 = No
edad	Age	Years old
peso	Weight	kg
talla	Height	m
temp	Temperature	Centigrades
fc	Heart rate	Beats per minute
fr	Respiratory rate	Cicles per minute
tas	Arterial pressure, systolic	mm Hg
tad	Arterial pressure, diastolic	mm Hg
score	Severity score	Score points (0-16)
ing_disnea	Short of breath at the moment of hospitalization	1 = Yes, 2 = No
sato2sin	Oxygen saturation	Percentage of Saturated hemoglobin
urea	Blood urea	mg/dL
bun	Blood urea nitrogen	mg/dL
creat	Serum creatinine	mg/dL
colesterol	Serum cholesterol	mg/dL
gluc	Blood glucose	mg/dL
hb	Hemoglobin concentration	mg/dL
leucos	White blood cells	number of cells/µL
plaq	Platelets	number of cells/µL
linfos	Lymphocytes	number of cells/µL
monos	Monocytes	number of cells/µL
eos	Eosinophils	number of cells/µL
basof	Basophils	number of cells/µL
neutros	Neutrophils	number of cells/µL
k	Potassium	mEq/L
na	Sodium	mEq/L
cl	Chloride	mEq/L
ca	Calcium	mEq/L
ph	pH	Units of pH
pao2	Arterial oxygen partial pressure	mEq/L
paco2	Arterial carbon dioxide partial pressure	mEq/L
hco3	Arterial bicarbonate	mEq/L
dhl	Serum lactate dehydrogenase	IU/L
alat	Serum Alanine aminotransferase	IU/L

Table 3: Description of the variables in the working database *(continued)*

Column name	Description	Codification/units
aat	Serum Aspartate aminotransferase	IU/L
dimd	D-dimer	D-dimer µg/mL
fecha	Date of consultation	
fechacov1	Date of covid-19 testing	
fechahosp	Date of hospitalization	
fechalta	Date of hospital discharge	
fechainisint	Date of appearance of first symptoms	
motivoegre	Vital status at the moment of discharge	2 = Alive, 3 = Dead

2 Methods/analysis

2.1 Data wrangling

I read the data from a `csv` file into an R object named `dbcovid`. The first thing I noticed is that date variables are of type `character`, thus I used the function `dmy` to transform the date variables into date type:

```
dbcovid <- dbcovid %>% mutate(across(starts_with('fecha'), dmy))
```

The laboratory data must be of `numeric` type, but several are as `character` in `dbcovid`, hence there must be typos. Table 4 shows these errors. To fix this problem, I use a `character` vector to put the names of the problematic variables, employ `regex` patterns and create a "catch-all" function, with the aid of the package `stringr`.

Table 4: Typos in numeric variables

	Typos in numeric variables
urea	13,9
creat	0,81
bun	4..5
plaq1	160 000
plaq2	203 000
ca	A
aat	BA
alat	NA

```
#Check errors in numeric variables
nomcadenas <- c('urea', 'creat', 'bun', 'plaq', 'ca', 'aat', 'alat')
patron <- '([^\\d\\.])|(\\.{2,})' #Anything but digits or one decimal point.

fpatron <- function(x){
  x[str_which(x, pattern = patron)]
}

errores <- apply(dbcovid[, nomcadenas],2,fpatron )

# Catch-all function to detect errors in numeric variables
```

```r
farreglar <- function(x){
  x = str_trim(x)
  x = case_when(
    str_detect(x, '\\s') ~ str_replace_all(x, '\\s', ''),
    str_detect(x, '[:alpha:]') ~ str_replace_all(x, '[:alpha:]',''),
    str_detect(x, ',|\\.{2,}') ~ str_replace_all(x,',|\\.{2,}', '.'),
    TRUE ~ x
  )
}

#Fix errors in numeric variables
dbcovid <- dbcovid %>% mutate(across(all_of(nomcadenas), farreglar))
#Check if it worked
arregerrores <- apply(dbcovid[, nomcadenas],2,fpatron )
arregerrores # no mistakes

# Transform numeric variables to numeric type and check the structure
dbcovid <- dbcovid %>% mutate(across(all_of(nomcadenas), as.numeric))
dbcovid %>% select(all_of(nomcadenas)) %>% str
```

Next, I use the dates to calculate duration of some events, I also compute other variables such as the body mass index and the presence of obesity. As can be seen in table 3, the codification of dichotomic variables is not uniform; I change them accordingly using the number 1 to indicate that the feature is present. Finally, I transform all categorical variables to `factor` class.

```r
#Create new variables with dates, and eliminate the date variables.
dbcovid <- dbcovid %>%
  mutate(bmi = peso/talla^2,
         dayshosp = as.numeric(fechalta - fechahosp),
         duration = as.numeric(fechalta - fechainisint),
         daysdelay = as.numeric(fechahosp - fechainisint),
         obesity = ifelse(bmi >= 30, 1, 2)) %>%
  select(-starts_with('fecha'))

dim(dbcovid)
names(dbcovid)

# Change codification of dichotomous categorical variables.

dicot <- c('motivoegre','has', 'tabaquismo', 'dm', 'renal', 'autoinmunidad',
           'ing_disnea', 'obesity')
dbcovid <- dbcovid %>%
  mutate(across(all_of(dicot), function(x) ifelse(x == 2, 0, 1)))
str(dbcovid)

# Transform categorical variables to factors.
nomcateg <- c('motivoegre','sexo', 'ocupacion', 'nivsoc', dicot[-1])
dbcovid <- dbcovid %>%
  mutate(across(all_of(nomcateg), as.factor))
str(dbcovid)
```

2.2 Data exploration

Due to the number of variables, I decide to focus the exploration of data on the difference in the outcome. I use these results as the first filter to decide which variables are going to be in the prediction models.

Collins, Gary S., Johannes B. Reitsma, Douglas G. Altman, and Karel G. M. Moons. 2015. "Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD): The TRIPOD Statement." *Journal of Clinical Epidemiology* 68 (2): 112–21. https://doi.org/10.1016/j.jclinepi.2014.11.010.

Sanchez Talanquer, Mariano. 2020. "La letalidad hospitalaria por covid-19 en México: desigualdades institucionales." https://datos.nexos.com.mx/?p=1625.

Secretaria de Salud. 2020. "Informe técnico Diario Covid-19 méxico," December. https://www.gob.mx/cms/uploads/attachment/file/601308/Comunicado_Tecnico_Diario_COVID-19_2020.12.17.pdf.