Long Assignment 2022/2023

Pedro G. Ferreira, Alipio Jorge

November 2022 (version 14.11.2022)

Objectives

The main goal of this work is to analyze a database of medical records in Mexico and to use Data Science and Machine Learning and to predict high-risk Covid-19 patients.

Dataset

The data is extracted from a database of confirmed and suspected COVID-19 infections in Mexico, constituting the official COVID-19 data compiled. It was made publicly available by the Mexican Federal Government:

- https://www.gob.mx/salud/documentos/datos-abiertos-bases-historicas-direccion-general-de-epidemiologia
- https://www.gob.mx/cms/uploads/attachment/file/753710/Cierre_Datos_abiertos_hist_ricos_2020.pdf

The goal is to determine if a patient that has been exposed to SARS-CoV-2 virus will have a positive outcome or a worst outcome and is more likely to die than to survive (see variable `TIPO_PACIENTE`).

The data analyzed corresponds to the cases from the period of 2020-04-12 to 2021-01-31 in a total of 3,779,640 cases. You are advised to first work with small samples in order to explore models and hyperparameters.

The dataset contains a total of 40 features (labels in Spanish) and more details can be found here in the paper:

PLoS One. 2021 Sep 20;16(9):e0257234. doi: 10.1371/journal.pone.0257234. eCollection 2021.
`Identification of high-risk COVID-19 patients using machine learning`
Mario A Quiroz-Juárez , Armando Torres-Gómez, Irma Hoyo-Ulloa, Roberto de J León-Montiel, Alfred B U'Ren * https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0257234

You are given one unique dataset. You will need to create the train and test datasets:

1. shuffle the dataset entries.
2. create a first division of the data into train (Tr) and test (Ts) using a proportion of 80/20 or 70/30.
3. Use the training data for calibration of your models (e.g. using Cross validation). You can use sampling, cross validation and other approaches. Use smaller samples to test ideas and reduce computation time. Use larger samples to obtain more reliable estimates of model performance.
4. Reserve the Ts dataset to evaluate the final models after all modeling decisions are made.

In the work of Quiroz-Juárez et al., the authors stratify the data into patients within four different clinical stages (see section Results of the paper). While this may not be strictly necessary, it is one possible approach to use. We do recommend starting with a predictive model without stratification.

Note that in this case, NA values are described by 97, 98 and 99 (eventually need recoding in Python):

- Don't know - 97
- Answer refused - 98
- Not applicable - 99

Guidelines

This data science problem should be approached by following the CRISP-DM methodology (http://jbusse.de/2019_ws_dsci/crisp-dm_phases-tasks-outputs.html). You have to understand the business problem, propose success criteria and see how it can be translated into a machine learning problem. Then you look at the characteristics of the data and you perform the required explorations, visualizations and transformations. Next step is to identify insights, develop predictive models and to evaluate them in order to validate if they are helpful in the business problems. During the whole process take notes, always identify the questions you want to answer and think before you act: "why is this plot or this transformation useful". You can perform some operations just for the sake of training but you should be aware of that.

The result is a **report** in the form of a **notebook** with clear explanatory text and code that works showing results. The report should be clear, as concise as possible and it should be easy to read and to follow. You will be telling the story of your approach to this problem, so it should have a good narrative flow. Always explain what you are doing, why you are doing it, what are the results and what do you take from those results.

Suggested structure

A report containing:

1. Business understanding
    - Give your view of the business problem following the CRISP-DM list of outputs when adequate.
2. Data Understanding
    - Looking at the raw data, describe variables according to their types: interval-scaled, binary, nominal, ordinal, ratio-scaled. Be aware that there are specific methods suitable to each type of variable.
    - Perform a preliminary analysis (summaries, spread measures, histograms, boxplots, density). These are interesting to be applied to the raw data to "uncover' ' inconsistencies, outliers, duplicates etc.
    - Perform bivariate analysis (correlations, regression)
    - Provide any insights about the data and the problem that you may have found.
3. Data Preparation
    - List of main changes that can need to be performed to the raw data, including feature selection.
    - Describe the potentially useful ones and their results in terms of data.
4. Modeling: consider the balanced and the non-balanced versions of the dataset as 2 separate problems. Firdt work with the balanced data and then with the non-balanced data. Try each of the methods below, select hyper parameters using default values and empirical analysis. Separate a test set and use cross-validation on the rest of the examples. Visualize models when possible, visualize results, produce aggregating tables with good insightful summaries of the results, and whatever other tools you may find useful.
    - Nearest neighbor
    - Bayesian Classifier
    - Decision Trees
    - Tree ensembles
    - Support Vector Machines
    - Neural Network Classifier
    - Comparison
5. Evaluation and Main Conclusions

- What is the best model and the recommended data science procedure for the business?
- What do you think that the business can gain from your data science effort?
- What are the lessons learnt?
- What is your summary of the achieved results?

To submit:

- a fully operational Rmd document or a Jupyter notebook with the selected experiments as clear and concise as possible. Avoid output dumps. Recall that the report is going to be evaluated by your very busy professors and that they may have to skip many pages if your report is too long. Always highlight your best results. Please note:
  - The objectives for each experiment and plot should be clear so that the reader understands why it is worth to read a particular part.
  - The conclusion should be a short high level account of what was observed.
  - It is **not necessary to describe the methods** (unless requested, but you should know their concepts and how they work). It is more important to point out the differences in the methods and the reasons for the results in terms of methods characteristics.
- A 5 minute video (or link to a video), per element of the group with a recorded presentation of the respective part of the work. The presentations of the group, when combined, describe the whole of the group's work.
- The project slide presentation.

Evaluation

- This assignment is worth the values described in sigarra, according to the course you are following.
- Components
  - Report 30%
    - Narrative 10%
    - Writing style 10%
    - Presentation 10%
  - Technical 70%
    - Diversity of the results for the experiments 20%
    - Correctness 30%
    - Challenge performance 10%
    - Conclusions 10%

Groups

Assignments are submitted by groups of 1 to 3 students. Different elements may have different grades. Other group sizes will not be considered.

It is advisable that the students from the same group perform overlapping work and only after that, exchange ideas with each other. Group work is important for learning from other people.

Submissions

Formal final deadline is **January 12th 2023**, to be submitted in moodle, and only in moodle. Submissions after that date will be multiplied by a monotonously decreasing factor that starts in 1.

- Checkpoints:
  - In the classes of **15th and 16th of December** each group should present an update with the status of the project. You will have 5 minutes for this presentation, where you can show your best results and list your main difficulties.
  - Intermediate submission on **December 16**: html version of the notebook with the first 2 CRISP-DM phases and part of the 3rd.

Ethical principles

When submitting, students commit themselves to follow strong ethical principles. All the work must be done by the elements of the group alone. All members of the group will be involved with the whole of the work. All the materials used and consulted must be credited in the work.