Statistical Machine Learning for Risk and Actuarial Applications
Assignment: Hospitalised COVID-19 Patients in Brazil
T2 2024

Due times:

Part one: Friday 28th June 2024 11:55 am (sharp)

Part two: Friday 26th July 2024 11:55 am (sharp)

1 Skills Developed

This assignment provides you with an opportunity to apply techniques you have learnt in the course to a business task involving data. In addition, your skills in understanding and applying advanced research works (including the text and any additional reference material you consider) will be developed via this assignment. Communication of the results of your investigations and analysis is also an important skill developed.

2 Task

You work as an actuary for an international insurance company which has some businesses in Brazil. Your role is to assess the mortality and health risk of the Brazilian private health insurance portfolio.

Your boss has just emailed you a set of Brazilian COVID-19 data along with a study by Hojo de Souza et al. (2021) about mortality and survival of hospitalised COVID-19 patients in Brazil. The study uses older data covering the period from February 26th to August 10th, 2020, but you also have access to data covering 1st January 2021 to 31st December 2021. Interestingly, this new dataset called `CovidHospDataBrasil.csv` has information on the vaccination status of the patients, which was previously unavailable. To inform decision making, your boss has asked you to analyse the new set of data. They want the analysis you deliver to be done in two parts, in particular you boss has asked your results to be done as:

2.1 Part one - The draft (30 marks)

1. Perform an exploratory data analysis of the profile of hospitalised COVID-19 patients. Provide detailed information (by graphs and statistics) on the patients' health characteristics and how they influence whether the patient was observed to die to COVID-19.
2. Fit an appropriate linear regression model (a model belonging to the GLM class) to model the relationship between patients' health characteristics and their hospital treatment, and those patients that ended up dying from COVID-19 (i.e., set `covidDeath` as your target variable). Demonstrate that your model is an improvement over a naive implementation of a full model that includes all variables. Demonstrate this through the use of indirect methods or through other facets of modelling, such as model interpretability or applicability.

2.2 Part two - Final report (70 marks)

1. Build a model to analyse the relationship between patients' health characteristics and their hospital treatment, and those patients that ended up dying from COVID-19 (i.e., set `covidDeath` as your target variable). In this section you are required to:

- Use a model from this course.
- Use at least one method of direct comparison between models.
- Fit at least one regularised or tree method.
- Consider a spline / polynomial implementation.

2. Your boss has asked you to build a predictive model to determine the likelihood of a newly admitted patient dying from COVID-19 based solely on the information available at the time of their admission. This approach aims to provide actionable insights for healthcare professionals (to whom your insurance company speaks to frequently) to prioritise care and allocate resources effectively.

See more details on the tasks below.

3 Additional information and mark allocation

3.1 Data

For the assignment you have access to a sample of hospitalised COVID-19 patients from the Brazilian Ministry of Health Database (SIVEP-Gripe). You have access to a pre-processed file "CovidHospDataBrasil.csv". This file contains data on 157209 individuals who were admitted to hospital due to COVID-19 between 1st January 2021 and 31st December 2021. Among other things, for each individual the dataset contains information on their clinical symptoms (fever, cough, sore throat, etc.) and pre-existing comorbidities (cardiovascular disease, asthma, diabetes, etc). The variables are split into two components, the **Variables known upon admission**, which are the variables recorded for a patient upon admission to the hospital, and **Variables known upon death or discharge**, which are the variables that are determined upon further testing or observation of the patient once admitted.[1] The data is described below:

Variables known upon admission

Variable	Description
`dateBirth`	Date of birth of the patient.
`age`	Age nearest birthday at the date when the patient was admitted to hospital.
`sex`	Sex of the individual taking the values `F` and `M` for females and males, respectively.
`vaccine`	`TRUE` if the patient had received at least one dose of COVID-19 vaccine or `FALSE` otherwise.
`dateHosp`	Date in which the patient was hospitalised.
`fever`	`TRUE` if the patient had fever or `FALSE` otherwise.
`cough`	`TRUE` if the patient had a cough or `FALSE` otherwise.
`sorethroat`	`TRUE` if the patient had a sore throat or `FALSE` otherwise.
`dyspnoea`	`TRUE` if the patient had dyspnoea or `FALSE` otherwise.
`oxygensat`	`TRUE` if the patient had blood oxygen saturation $< 95\%$ or `FALSE` otherwise.
`diarrhea`	`TRUE` if the patient had diarrhea or `FALSE` otherwise.
`vomit`	`TRUE` if the patient developed vomiting symptoms or `FALSE` otherwise.
`hematologic`	`TRUE` if the patient had any hematologic diseases or `FALSE` otherwise.
`downsyn`	`TRUE` if the patient had Down's syndrome or `FALSE` otherwise.
`asthma`	`TRUE` if the patient had asthma or `FALSE` otherwise.
`diabetes`	`TRUE` if the patient had diabetes or `FALSE` otherwise.
`neurological`	`TRUE` if the patient had any neurological diseases or `FALSE` otherwise.
`pneumopathy`	`TRUE` if the patient had a pneumopathy or `FALSE` otherwise.
`obesity`	`TRUE` if the patient had obesity or `FALSE` otherwise.

Variables known upon death or discharge.

[1] Note, these variables were split for educational purposes. The original dataset did not determine which health markers would be known upon admission or only after admission.

Variable	Description
`dateEndObs`	Date of the end of observation. This can be either the date of death or the date the patient was discharged from hospital.
`covidDeath`	`TRUE` if the patient died from COVID-19 or `FALSE` if the patient recovered from COVID and was discharged.
`icu`	`TRUE` if the patient was admitted to an intensive care unit (ICU) or `FALSE` otherwise.
`dateAdmIcu`	Date in which the patient was admitted to ICU. It can be greater or equal to `dateHosp`.
`dateDisIcu`	Date in which the patient was discharged from ICU. It can be smaller or equal to `dateEndObs`.
`respdistress`	`TRUE` if the patient developed acute respiratory distress syndrome (ARDS) or `FALSE` otherwise.
`cardio`	`TRUE` if the patient had a cardiovascular disease or `FALSE` otherwise.
`hepatic`	`TRUE` if the patient had any liver diseases or `FALSE` otherwise.
`immuno`	`TRUE` if the patient had any immunodeficiencies or `FALSE` otherwise.
`renal`	`TRUE` if the patient had any renal diseases or `FALSE` otherwise.

Target variable

Variable	Description
`covidDeath`	`TRUE` if the patient died from COVID-19 or `FALSE` if the patient recovered from COVID and was discharged.

3.2 Part one (30 Marks)

The primary purpose of splitting the assignment into two parts reduce the burden of one major deadline into two manageable pieces. Part two builds directly on part one. Part one consists of the preliminary investigations into the COVID-19 hospital dataset.

Mark allocation for part one of the assignment can be found in the rubric attached, and also refer to the information below for more details on the tasks.

3.2.1 Exploratory data analysis of the patients' health characteristics and hospital treatment (10 Marks)

For this part you should do exploratory data analysis (EDA) of the dataset to summarise the conditions of the patients that ended up dying from COVID-19. Your presentation of the different EDA summary metrics should be accompanied by a discussion of the insights you get from the summary metrics.

It is advised for you to consider multiple graphs and metrics, and to select which graphs and metrics that provide interesting insights into the dataset. These insights, can guide you in the next stage when modelling the dataset.

The StoryWall formative activity for Week 3 also involves EDA of this dataset. You can use the answers and discussions from that activity as an starting point for this task.

3.2.2 Fit an appropriate linear model for the modelling of COVID-19 deaths based on patients' health characteristics (15 Marks)

For this part you are required to fit an appropriate linear model of your choice to model the characteristics of patients that eventuate in death from COVID-19. In this section you are required to conduct the following:

- Fit a linear model of your choice from the GLM family. You are free to take any GLM model to achieve this task (but some are more appropriate than others).

- You must provide proof that your model is an improvement over a naive implementation of taking all variables. This proof can be in the form of indirect methods (AIC, BIC, etc.) and through other relevant aspects of modelling, such as interpretability.
- Provide a discussion on the insights you get from your model.

Provide details about the construction of the model in the technical appendix.

3.2.3 Presentation and communication (5 Marks)

Communication of quantitative results in a concise and easy-to-read manner is a skill that is vital in practice. As such, marks will be given for the presentation of your results. In order to maximise your presentation marks you may wish to consider issues such as: table size/readability, figure axis/formatting, ease of reading, grammar/spelling, and report structure. You may also wish to consider the use of executive summaries and appendixes, where appropriate. Provide sufficient details in the main body of the reader so that they can judge what you are doing, using appendices for non-essential but useful results as necessary.

In writing the report, you can assume that the client and your manager are both familiar with the details of statistical learning (as if you were presenting to a fellow classmate). Note that sufficient detail must be provided (in either the report body and/or appendices) so that the reviewer can follow all the steps and derivations required in your work.

For part one there is **maximum page limit of 3 pages** (including tables and graphs but excluding references) is applicable to the main body of the report.[2] You should also consider the rubric for the presentation component. There is no limit to the size of the appendix. Furthermore your answer should satisfy the following formatting requirements: (i) font: Times, 12 pt or equivalent size and (ii) margins: all four of at least 2 cm.

3.3 Part two (70 Marks)

Part two of the assignment is for you to flex your statistical learning knowledge and skillset to the fullest! You will be expected to provide more compelling evidence and justifications in part two compared to part one of the assignment. **Part two of the assignment should be self-contained.** Do not reference your part one of the assignment unless you have provided the details in the part two of your assignment. You are not required to have part one attached in the submission of part two.

Note that, as in any consulting exercise, there are many alternative valid approaches that can be used. You can choose how to perform the tasks as long as they are justifiable and justified. What is important is the rationale for your chosen methods, and the associated insights and recommendations.

You may also wish to engage in extra research beyond what is covered in the course – please feel free to do so. Although the marks for each component of the assignment are capped, innovations will be encouraged and will potentially offset issues if present. Note however that it is possible to attain full marks without significant innovation.

Mark allocation for part two of the assignment can be found in the rubric attached, and also refer to the information below for more details on the tasks.

3.3.1 Modelling the relationship between patient characteristics and COVID-19 deaths (30 Marks)

For this part you should build a model to understand and interpret the relationship between fatal accidents and the different variables available in the `CovidHospDataBrasil.csv` dataset. For this component you are restricted to **models only from this course**.

This component is similar to Q2. of Part one of the assignment. However, here you are required to demonstrate a higher level of competency and conduct the following

[2]Please kindly note that this is a maximum - you should feel free to use less pages if it is sufficient!

- Use direct methods of comparisons, including CV or k-folds CV.
- Use at least one regularisation method or tree method.
- Consider a spline / polynomial implementation.

You should provide in the main report the results a detailed analysis using your selected model, along with justification of why the particular model was chosen. Your analysis should also be accompanied by a discussion of the insights you get from the model. Provide details about the construction of the model in the technical appendix.

3.3.2 Predicting at high-risk individuals COVID-19 admission (20 Marks)

Your boss has proposed creating a predictive model to determine the likelihood of a patient's death based solely on the information recorded upon hospital admission. This model aims to assist medical practitioners, with whom your company frequently collaborates, by providing valuable insights to identify high-risk individuals.

For this part you should develop a predictive model based only on the information that hospitals have for patients that are only admitted, i.e., from the **Variables known upon admission** variables. Given an individual has been admitted to hospital, based on their health characteristics, predict whether an individual is at high risk of dying, so that medical attention can be prioritised for these individuals.

Using the information you have in **Variables known upon admission** (see section on data for details), develop predictive models using various methods such as (but not limited to): logistic regression, k-nearest neighbours, logistic regression with lasso and ridge, classification trees and their extensions. Provide the results and analysis associated with each of these methods in the technical appendix; this should include discussion on the choice of the tuning parameter(s). A very brief summary of each approach should also be included in the main body.

Note that you should also provide in the main report the results and a detailed analysis using your selected predictive model, along with justification of why the particular model was chosen. **Note: You are not restricted to models from the course for this component.** If you wish to fit more state-of-the-art models you may do so, but the requirements of providing reasoning and details on your model choice are still the same. Fitting a model outside of the scope of this course will not guarantee a higher mark. **Extensive details of your model should be placed in the technical appendix if you wish to do this.**

You are also required to provide a description of the health characteristics of patients that are more likely to result in death from COVID-19. This description should be clearly supposed by your predictive modelling results.

3.3.3 Marks for predictive performance (10 marks)

The quality of your predictions on the evaluation data will have a (minor) impact on your mark. More specifically, 10 out of the 100 marks across the whole assignment will be associated to the accuracy of your predictions.

The marks you will get for the accuracy criterion will be based on the criterion used in the Kaggle competition setup, which will be set up in due time. Marking scheme will be announced at a later date.

3.3.4 Presentation Format and Communication (10 Marks)

Similar to part one of the assignment, marks will be given for the presentation of your results for part two. Refer to the communication section in part one for further details. **Part two of the assignment should be self-contained.** Do not reference your part one of the assignment unless you have provided the details in the part two of your assignment. You are not required to have part one attached in the submission of part two.

For part two there is a **maximum page limit of 5 pages** (including tables and graphs but excluding references) applicable to the main body of the report. You should also consider the rubric for the presentation component. There is no limit to the size of the appendix. Furthermore your answer should satisfy the

following formatting requirements: (i) font: Times, 12 pt or equivalent size and (ii) margins: all four of at least 2 cm.

3.4 Software

You may choose which software language to use (e.g., `R`, `Python`, or other), however, nearly every function you will be required to use for this task is available in `R`. Note also that code enabling you to perform most of the modelling can be found in the learning activities of the course. If you use any simplifying assumptions in your modelling, they must be clearly identified and justified.

3.5 Assignment submission procedure

3.5.1 Turnitin submission

Your assignment report must be uploaded as a **unique document**. As long as the due date is still future, you can resubmit your work; the previous version of your assignment will be replaced by the new version.

Assignments must be submitted via the Turnitin submission box that is available on the course Moodle website. Turnitin reports on any similarities between their own cohort's assignments, and also with regard to other sources (such as the internet or all assignments submitted all around the world via Turnitin). More information is available at the UNSW Turnitin page. Please read this page, as we will assume that you are familiar with its content.

Please **also attach any programming code and/or sample spreadsheet output** used in your analysis as a separate file in the dedicated "code_sample" Moodle assignment box on the course webpage. These will be referred to by the marker only if needed, and in particular the **main assignment (with appendix) should be self contained**.

3.5.2 Late submission

Please note that it is School policy that late submission of assignments will incur in a penalty.

When an assessment item had to be submitted by a pre-specified submission date and time and was submitted late, the School of Risk and Actuarial Studies will apply the following policy. Late submission will incur a penalty of 5% per day or part thereof (including weekends) from the due date and time. An assessment will not be accepted after 5 days (120 hours) of the original deadline unless special consideration has been approved. An assignment is considered late if the requested format, such as hard copy or electronic copy, has not been submitted on time or where the 'wrong' assignment has been submitted. **Students who are late are still required to upload documents to the appropriate submission boxes.** The date and time of the last Moodle submission determines the submission time for the purposes of calculating the penalty.

You need to check your document once it is submitted (check it on-screen). **We will not mark assignments that cannot be read on screen.**

Students are reminded of the risk that technical issues may delay or even prevent their submission (such as internet connection and/or computer breakdowns). Students should then consider either submitting their assignment from the university computer rooms or **allow enough time (at least 24 hours is recommended) between their submission and the due time. No paper copy will be either accepted or graded.**

3.5.3 Plagiarism awareness

Students are reminded that the work they submit must be their own. While we have no problem with students working together on the assignment problems, the material students submit for assessment must be their own.

Students should make sure they understand what plagiarism is — cases of plagiarism have a very high probability of being discovered. For issues of collective work, having different persons marking the assignment

does not decrease this probability.

3.5.4 Generative AI policy

You are allowed to use Generative AI to help you with editing, planning, idea generation, or coding. However, please include an Appendix in the report titled **"Generative AI usage"** explaining what you used AI for and, if applicable, outlining what prompts you used. If **you did not use Generative AI write in this Appendix that generative AI was not used**.

References

Hojo de Souza, Fernanda Sumika, Natália Satchiko Hojo-Souza, Ben Dêivide de Oliveira Batista, Cristiano Maciel da Silva, and Daniel Ludovico Guidoni. 2021. "On the analysis of mortality risk factors for hospitalized COVID-19 patients: A data-driven study using the major Brazilian database." *PLoS ONE* 16 (3 March). https://doi.org/10.1371/journal.pone.0248580.