Assignment for Module 5

Review of:

Deciphering C-section Rate Disparities: Analyzing Swiss Hospital Traits for Insights

Author: Gaëlle Marti
Contact: gaelle.marti@protonmail.com

Submitted by: Sebastian Schaller
Contact: sebastian.schaller@unibe.ch

Date: 27.12.2023

Introduction

As part of the Certificate of Advanced Studies in Applied Data Sciences (CAS ADS), Gaëlle Marti, Simon Rime, and Lenja Flütsch are investigating cesarean rates in Swiss hospitals. They started the project with Module 1, where they presented an outline of the aim and concept of their study. They developed it further during Module 2, where they performed classical statistical analyses of the data, and Module 3, where they performed advanced data analyses with machine learning techniques. In this review, I will give general feedback on the whole project and Gaëlle Marti's contribution to Module 3. In which she applied decision tree, random forest, and hierarchical clustering methods for advanced data analysis.

Project Description

The project aims to investigate and possibly explain the observed differences in the cesarean rates in Swiss hospitals based on publicly available data annually published by the Swiss Federal Office of Public Health from 2015 to 2021. The final dataset consists of 17 features, such as the number of beds, size of the medical staff, etc., and has 496 rows, each representing a hospital in a given year between 2015 and 2021. Therefore, the combined dataset contains data from the same hospitals from different years. The existence of a significant difference in the Caesarean section rate between the Latin and German-speaking parts of Switzerland was statistically proven and demonstrated in Module 2. However, the exact reason for this is unclear. Furthermore, this likely culture-related aspect cannot explain the observed internal differences in the two populations. In Module 3, various supervised and unsupervised machine learning and advanced data analysis methods are applied to investigate further the relationship between the traits and the cesarean section rate, e.g., linear and logistic regression, random forest and decision tree models, principal component analysis (PCA), and hierarchical clustering.

Strength of the project

The project is well-structured and well-explained on a conceptual level, as outlined in the Module 1 report, which was further developed and improved during the work for Modules 2 and 3. For the project, solid workflows for collecting publicly available data via web scraping, reliable criteria-based quality control, and simple but efficient data processing were designed, tested, and presented during Modules 1, 2, and 3. The formulated research question is relevant to several domains, such as medicine, health policy, and ethics. It is supported by solid domain knowledge and a literature review. Furthermore, its significance was tested, proven, and presented during the statistical analysis of the data set during Module 2.

A sound approach to the study and evaluation of the cleaned dataset was developed during Module 1 and mainly tested during Module 3. A wide range of supervised and unsupervised machine learning techniques, such as linear regression, random forest, decision tree, principal component analysis (PCA), and hierarchical clustering, were applied and evaluated individually and in combination. Using different

techniques for data analysis enhances the visualized aspects of the data sets and strengthens the reliability of the results. Each method can highlight a different aspect of the data set, which the others may overlook. The applied techniques and models are suitable for a wide range of data analysis without being too resource-consuming.

The revised part of Module 3 contributes by applying a random forest and a decision tree model to the dataset. These two models evaluate the importance of features and provide a reduced feature catalog without losing significant information. The results of the decision tree and random forest models are visualized and presented clearly and understandably. In addition, hierarchical clustering was applied to the entire dataset with all 468 features and to the reduced dataset with the 17 most significant features. The results of both clusterings are visualized using the corresponding clustering dendrograms and box plots. The selected number of clusters used is determined based on the silhouette score. The individual Jupiter notebooks are structured so that they can be followed by an external person with a similar level of knowledge as the author.

Suggestions for improvement

The project has a solid concept and an interesting, well-chosen, and tested research question. The tools for data analysis and answering the formulated questions are reliable and appropriate. However, I strongly recommend not deleting the hospital names from the dataset during data processing and analysis. I suggest introducing a unique identifier for each row of data, consisting of a combination of the hospital name and the year from which the data originated or a code that allows each row of the data to be uniquely identified. Otherwise, there is no way to correctly identify and examine individual data points, making it impossible to distinguish between true outliers and errors in the data. This would also improve the data quality and ensure better quality control and robustness of the analysis. In addition, I am not sure that combining data from multiple years into one set is the best approach for time series analysis. Since not all hospitals are equally represented in the dataset due to missing or erroneous data or because they were added during the survey, this unbalanced representation of individual hospitals may bias the data. It may be helpful to analyze each year's data separately or to use only those hospitals with valid data over the entire time series. In both cases, some of the bias from the over-representation of the data can be reduced, and a proper time series analysis of the entire dataset and individual hospitals can be performed, which may allow for further detailed analysis of the data.

Another point I recommend to rethink is the project's documentation in general. Using different Jupiter notebooks for each method in combination with a shared Google-Drive is a valid solution for working together on a simple project. However, even if the individual notebooks/scripts are well documented and structured, the task's fragmentation may make it hard to keep track of and understand the dependencies of the different notebooks/scripts and how they work together. It is possible to lose track and make it difficult to fully understand the code, especially when only parts of the project are provided, as was the case for this review. This problem could be improved by using an advanced version control

system and online repository, such as Git and Github, or at least including some additional explanation of how the various code fragments work together. This would also improve the reproducibility of the results, following the FAIR principle[1].

The general feedback covers most of the possible problems or concerns in the reviewed part of Module 3. The individual notebooks are mostly clear, except for minor inconsistencies in the level of detail in plotting the results and commenting out the codes. Improving these would smoothen out the notebook/scripts nicely without much effort. A more significant issue is documenting and preserving the exact input and output of decision tree and random forest models since they produce different results each time the code is run, even using the same training and test sets, making it impossible to reproduce the exact results. This makes it difficult to justify decisions based on the output of the models when they are incompletely documented. I would recommend diving into the scikit-learn documentation of the decision tree[2] and random forest[3] models and finding a way to preserve all the parameters and metadata needed to reproduce the exact input and output. The last major problem is probably caused by a lack of overview between the different notebooks/scripts. How and why the reduced features were selected for hierarchical clustering was unclear. I assumed it was based on the importance of the feature extracted from the random forest and/or decision tree models. Since different datasets are used for different tasks, it is impossible to understand this without additional information. I recommend being more consistent in this as well.

In a further step, I think it would be worth investing more time into thinking about the meaning and correlation of the individual features. There may be additional features to consider, such as the number of delivery rooms and beds in the maternity ward or the number of midwives on the medical staff.

Conclusion

The project is well-structured, follows a well-planned workflow, and focuses on a relevant and interesting topic with a clear and well-defined research question. Some more work needs to be invested in the documentation to ensure the reproducibility of the results according to the FIAR principle. In addition, I would recommend investing some more time in thinking about the data, such as what additional features might be helpful for the analyses and increasing domain knowledge. Overall, the project is on a good track, and I am excited to see the study's final results.

Suggested readings

[1] M. D. Wilkinson *et al.*, "The FAIR Guiding Principles for scientific data management and stewardship," *Sci Data*, vol. 3, no. 1, Art. no. 1, Mar. 2016, doi: 10.1038/sdata.2016.18.

[2] "1.10. Decision Trees," scikit-learn. Accessed: Dec. 27, 2023. [Online]. Available: https://scikit-learn/stable/modules/tree.html

[3] "1.11. Ensembles: Gradient boosting, random forests, bagging, voting, stacking," scikit-learn. Accessed: Dec. 27, 2023. [Online]. Available: https://scikit-learn/stable/modules/ensemble.html