Healthcare Analytics Group Project (updated 15 July 2024)

Real-world data (RWD) is potentially a rich and valuable source of information that can be utilized to improve the delivery of health services. From the data analytics perspective, RWD derived from electronic medical records (EMR) contains a wealth of information that can be mined for a variety of purposes, including producing real-world evidence (RWE) for regulatory purposes, risk factor analysis to guide policy decisions and the development of risk prediction models

For this group project, EMR data from SGH over the period from 1 Jan 2012 to 31 Oct 2016 are retrospectively extracted and provided. This data contains information on patient demographics, comorbidities, laboratory results, surgical priority and surgical risk amongst others. Outcome measures are days between surgery and death, mortality event, death within 30 days after surgery and ICU admission. The dataset contains nearly 80,000 surgical cases. The background of the dataset can be found in Chan et al. (2018)[1]. Group project repository can be found here:

https://github.com/ISSS623-AHA/ISSS623_2024/tree/main/Group_Project/Group_Project-SGH

Aims

You are required to utilize this dataset for the following:

(1) Risk factor analysis – evaluate the risk factors associated with the outcomes
(2) Predictive modelling – develop validated predictive models for the prediction of various outcomes

Specific Tasks

You should show competencies in performing the following tasks:

(1) Data wrangling and preparation. Examples of competencies in this area include:
    - Missing data imputation methods (try different imputation methods and compare the robustness of results
    - Data quality assessment (visualize and summarize the data used in the models
    - Dealing with bias, imbalanced data, etc(improve on them if possible)
(2) Risk Factor Analysis. Examples of competencies in this area include:
    - Univariate and multivariate analysis
    - Model quality evaluation
    - Model improvement
    - Model analysis and interpretation of results
(3) Predictive Model Development. Examples of competencies in this area include:
    - Test out different models (machine learning/statistical models)
    - Test out different validation metrics (include the most common ones)
    - Hyperparameter tuning (find the best hyperparameters)
    - Improve on the Training/Validation/Testing procedures (e.g., dealing with missing values, outliers, imbalanced data)

For all the above tasks, you should also ensure the following:

(1) Always ensure that the information is available at the time of prediction
(2) Understand the data assumptions as described in Chan et al. (2018)[1]
(3) Rigour and accuracy of the model prediction and validation processes
(4) Comparison across different models is mandatory

(5) Develop relevant visualization to support your model development process
(6) Submit a presentation of 30 slides of less (additional information can be put into the Annex)

Additional Notes:

(1) This is not a Kaggle Competition nor a coding competition (you can use other software if they are more effective)
(2) Scores will not be based solely on prediction accuracy
(3) Document your assumptions and limitations (follow standard reporting guidelines)
(4) Focus on the concepts to improve the analysis
(5) Highlight all limitations. Reporting correctly following key aspects of recommended reporting guidelines (e.g., STROBE[2]) will be recognized
(6) Coding style will NOT be evaluated. But the codes have to be runnable to reproduce the results
(7) No presentation is needed. A presentation slide deck must be submitted for this project.

Suggested Presentation Structure:

1. Title/ Abstract
2. Introduction/ Background
   - Literature Review
   - Data understanding
   - Project Management
3. Data and Methods:
   - Data Wrangling
   - Feature Engineering
   - Descriptive Statistics
   - Model Development and Validation
4. Results and Discussion
   - Data visualization (relevant ones)
   - Comparison of models
5. Limitations and Assumptions
6. Future work
7. Conclusion
8. References
9. Appendix

Key deliverables:

(1) Presentation deck that is 30 slides or less (additional information can be put into the Annex).
(2) Codes (in Python/R, or anything that the group is familiar with), relevant visualizations across the healthcare analytics process
(3) For the Testing/Validation of model, there is no target accuracy level, but group is able critique the performance of the model and the process of evaluating model performance

Challenge Notes

Use of real-world datasets and modelling real-world problems may not lead to a single correct answer. Hence, do not attempt to come up with a perfect solution. What we hope to see in the group submission is the ability to work out a reasonable scope and carry out what have been proposed, that should be sufficiently logical and reasonable.

Groups are reminded to not try to perfect the models or analysis given the limited timeline. We hope students can demonstrate competencies in the various tasks with the scope of the activities appropriately planned and executed.

References

1. Chan DXH, Sim YE, Chan YH, et al. Development of the Combined Assessment of Risk Encountered in Surgery (CARES) surgical risk calculator for prediction of postsurgical mortality and need for intensive care unit admission risk: a single-center retrospective study. *BMJ Open* 2018; 8: e019427.

2. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. *Journal of Clinical Epidemiology* 2008; 61: 344–349.