Exploration and Modeling of a MIMIC III Cohort

This project has two aspects: team and individual activities. The individual activities are graded, while the team activities prepare you for these.

Team activities.

With your team agree upon a disease or health condition for study utilizing the MIMIC III dataset, which you have been granted access to by Physionet. You may choose heart disease, diabetes, sepsis, respiratory illness, or some condition that interests you. Select a prediction problem, such as mortality or classification of patients who have this disease from those who do not.

As a team develop inclusion and exclusion criteria for cohort selection. For example, if you are interested in patients who developed a condition while in the hospital, you may set ICD-9 inclusion criteria for that condition, and an admission diagnosis as exclusion criteria. As a team create SQL or R code to select your cohort from the MIMIC III dataset. Starter code and other aides have been developed by Rohini, and are linked below.

With your team you may develop code to create logistic regression, SVM, decision tree and random forest predictive models for MIMIC III data, but only using the thromobocytopenia data cohort, extracted by Rohini, which is in the file store on Canvas. Once you have extracted your own cohort from MIMIC III using your team's inclusion and exclusion criteria, there is to be no viewing or sharing of code or plots with other students. All coding collaboration is to be done on the thrombocytopenia dataset as a sandbox (learning) environment.

Individual activities.

After the collaborative phase, you should have sample code to fit and evaluate the above models using MIMIC III data. You also should have a cohort of patients that have been selected from the MIMIC III dataset by your team. From here on you will work individually, without collaboration, to clean the data, perform exploratory analysis and model the data (including evaluation of models) so as to investigate and address the problem your group selected. You should develop 4 models: logistic regression, SVM, decision tree and random forest models. You may substitute elastic net for logistic regression and/or a boosting model for random forest. Evaluation of models should, at a minimum, include confusion matrix and ROC curves.

Create a report of your findings, either in R Markdown or in a Word document. Upload both your code and your report in Word or html. For your report, let your comments be brief interpretations of your visualizations and summary output.

Please be sure to individually (without collaboration) complete the following steps:

1. Perform data cleaning.
2. Display summary information on the data.
3. Create visualizations of the distributions of key variables by the response variable. (Ex: colored by mortality.
4. Create visualizations of a couple of relationships you find interesting between variables (ex: scatter plot colored by mortality).
5. Split your data into train and test sets.

6. Fit and evaluate a logistic regression or an elastic net model. Be sure to include regularization and evaluate with pseudo R^2 and AIC/BIC. Consider providing a plot to visualize relationships revealed by your model.
7. Fit and evaluate an SVM classifier, trying linear, poly and RBF kernels. Be sure to tune hyperparameters so as to avoid underfitting and overfitting.
8. Fit and evaluate a decision tree. Be sure to tune hyperparameters.
9. Fit a random forest or boosting model. Be sure to tune hyperparameters
10. Briefly summarize your findings, including a few sentences stating what your plots reveal and evaluating/comparing your models.

Resources for cohort selection:

View Rohini's recorded lecture on MIMIC III :

https://drive.google.com/file/d/1Gu29Y-rl5vncCYnldKR27i-xga7ciCk1/view?usp=sharing (Links to an external site.)Links to an external site.

Additional resources from Rohini:

Presentation slides:

https://docs.google.com/presentation/d/1GJ4r4egrAETWnX4m0mHjFFGxeJMuoY97sFmvbTFhO6c/edit#slide=id.p (Links to an external site.)Links to an external site.

Markdown file

https://drive.google.com/file/d/1lIHHABaYQ900SFPTo_wgxVvCU2fb3F8j/view

MIMIC III data download instructions

https://docs.google.com/document/d/1Rqz7_aMDNEP12UvaVgV37L2CsB-EZPg3KMtVMrQDo-8/edit