Learning Data Science with SAS® University Edition and JupyterLab

Brian Gaines

Think of a time when you needed to communicate the results of a data analysis

- Did it suffer from the “two program” problem?
  - SAS: code, results, graphs, etc.
  - Word/LaTeX: explanatory text and math notation
- Did it involve a lot of copy and paste?
  - Or input/output overhead?

Think of a time when you needed to communicate the results of a data analysis

Here, we estimate the logistic regression model

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 \text{sex} + \beta_2 \text{age} + \beta_3 \text{sex*age} + \beta_4 \text{pclass} + \beta_5 \text{fare} + \beta_6 \text{famSize}$$

by using the training set. The results include the Classification Table containing error estimates for different cut-off points, in addition to fit statistics for the training and test sets (see Appendix).

Based on the results, a cut-off point of 0.52 instead of 0.50 is used to make predictions on the test set to assess the model's performance on new data. In this situation, the resulting misclassification error rate is the same, and the model can correctly classify passengers about 77% of the time.

Analysis Variable : predError
Mean
0.2461538

What if I told you that there is a better way?

- Executable code
- Results
- Descriptive text
- Math notation

All in one document!

What if I told you that there is a better way?

Based on the Classification Table above, a cut-off point of 0.52 instead of 0.50 is used to make predictions on the test set to assess the model's performance. In this situation, the resulting misclassification error rate was the same, and the model is able to correctly classify passengers about 77% of the time.

[19]:
```
data work.titanicTestPred;
    set work.titanicTestPred;
    if P_1 > 0.52 then survivedPred = 1;
        else survivedPred = 0;
    if survivedPred = survived then predError = 0;
        else predError = 1;
run;
proc means data=titanicTestPred mean;
    var predError;
run;
```

[19]:

The SAS System

The MEANS Procedure

Analysis Variable : predError
Mean
0.2461538

What if I told you that there is a better way?

Based on the Classification Table above, a cut-off point of 0.52 instead of 0.50 is used to make predictions on the test set to assess the model's performance. In this situation, the resulting misclassification error rate was the same, and the model is able to correctly classify passengers about 77% of the time.

[19]:
```
data work.titanicTestPred;
    set work.titanicTestPred;
    if P_1 > 0.52 then survivedPred = 1;
        else survivedPred = 0;
    if survivedPred = survived then predError = 0;
        else predError = 1;
run;
proc means data=titanicTestPred mean;
    var predError;
run;
```

[19]:

The SAS System

The MEANS Procedure

Analysis Variable : predError
Mean
0.2461538

SAS resources for educators and learners

Academic Programs

- Joint Certificate Program
- Master's Degree Program
- Student Ambassador Program

SAS Software

- SAS® University Edition
- SAS® on Demand for Academics
- SAS® Viya® for Learners

Training

- On-site faculty and/or student workshops
- Free e-Learning
- Video Tutorials
- Training & Certification Discounts

Teaching Assets & Resources

- Teaching Kits
- Free Evaluation Book Copies
- Curriculum Consulting
- Online Support Communities

sas.com/academic

SAS resources for educators and learners

Academic Programs

- Joint Certificate Program
- Master's Degree Program
- Student Ambassador Program

SAS Software

- SAS® University Edition
- SAS® on Demand for Academics
- SAS® Viya® for Learners

Training

- On-site faculty and/or student workshops
- Free e-Learning
- Video Tutorials
- Training & Certification Discounts

Teaching Assets & Resources

- Teaching Kits
- Free Evaluation Book Copies
- Curriculum Consulting
- Online Support Communities

sas.com/academic

SAS University Edition is **free** SAS software for academic, non-commercial use

- Install locally or access via Amazon Web Services
  - Available on Windows, Linux, and macOS
- Includes several SAS products
  - Base SAS®, SAS/STAT®, SAS/IML®, some forecasting procedures

SAS University Edition includes two programming interfaces

- SAS Studio
  - Modern, browser-based SAS programming interface
  - Programming tools, code-generating tasks, code snippets
- JupyterLab
  - Open-source, browser-based programming interface
  - Successor to Jupyter Notebook
  - Jupyter notebooks: executable code, results, text
  - Can also be used with paid SAS license

SAS University Edition and JupyterLab facilitate learning important data science skills

- Computing
- Communication
- Collaboration
- Reproducibility

JupyterLab in action

Live demo

SAS University Edition and JupyterLab are a great combination for learning and teaching data science

- Free and easy to use
- Improves workflow
- Sharpens communication skills
- Facilitates collaboration
- Promotes reproducibility
- Popular in industry
- Interactive lecture notes for active learning

Thank you!

Brian Gaines
Brian.Gaines@sas.com
http://brgaines.github.io

sas.com