

DAT5 SF: HOMEWORK 4 “MIDTERM” – LOGISTIC REGRESSION

Assigned: Thursday, April 3

Due: Sunday, April 13 by 11:59PM

Submission Method: Submit via Schoology. Feel free to push to Github after the deadline has passed. Work independently on this midterm assignment.

This homework is a formally assessed “midterm” assignment. It will be evaluated on a 0-100 point scale.

The purpose of this homework is to gain experience implementing logistic regression, as well as to pull together much of what we have learned to date. This includes data wrangling, imputation, model evaluation (e.g. cross validation, ROC curves), visualization (e.g. matplotlib), and thoughtful interpretation of results.

DATA & CONTEXT

For this assignment, we will use the Titanic dataset that we explored in HW2. In this assignment, we will build a full logistic regression model to predict passenger survival.

SUBMITTING YOUR WORK

You may work in iPython notebook, via Python scripting, or both. Note that iPython notebook provides the ability to export Python scripts. Filenames for Python scripts should end in “.py”. To execute a Python script from the command line, simply type:

```
python <filename.py>
```

Please submit your work via Schoology.

HOMEWORK QUESTIONS

In addition to your code, please submit a 1-3 page report discussing your approach, results, conclusions, and next steps. Include visualizations. As part of your overall report, please address the following specific questions:

1. Several passengers are missing data points for age. Impute the missing values so that there are no “NaN” values for age as inputs to your model. Explain what value or values you used for the missing age data and why you used those values. (Hint: Think statistically here. Explain your thinking and choices.)
2. Create and run a logistic regression on the Titanic data to predict the survival of passengers.
 - a. Make sure your model results are reproducible by the instructor team, given any random aspects to your code / steps.
 - b. For your model output, include coefficient values, residuals, and standard errors.

- c. Explain which features are predictive for this logistic regression and, intuitively, why you think this may be so. Describe your thinking / logic in words. It is not sufficient to cite output statistics.
 - d. Make sure that you can push new data points through your model to get regression output. We will be providing a test set for this purpose.
3. Implement cross-validation for your logistic regression. Pick the number of folds. Explain your choice.
4. Create an ROC curve for your logistic regression by varying the threshold value for survival.
 - a. Plot this ROC curve visually.
 - b. What is the AUC for your model?
 - c. Explain why the model has achieved this level of accuracy/precision.
 - d. Explain how you could improve these metrics, as potential next steps.
 - e. Armed with this knowledge, what threshold value would you use? Why? Describe your thinking.