David Kinney DSC 550

Professor Deitenbeck Winter 2019

8.4 Reflection: Week 8 Reflection

I managed to get myself confused a bit in regards to keeping classifiers and regressors separated. I attribute it to the fact that I’m focusing on classifiers for the Titanic exercise and regressors for my case study. In fact, I’ve gotten a little obsessed with the Titanic study. While I’ve been a member of Kaggle for about two years, “member” was about it. The last couple weeks I’ve been really trying to get out of my Data Science comfort zone, so I entered the NOOB competition; which, it just so happens, is the Titanic study.

While the exercise we’ve been going through in class utilizes a Logistic Regression model, I’ve been tinkering with a Random Forest Classifier. In addition, for my Use Case, the end goal will be to predict a numeric value—Life Satisfaction Rating—based on all numeric features. So, for that I am using a Random Forest *Regressor.*

Where I am getting turned around is not so much the model *selection*, but the model *evaluation*. As an example, I wrote down all the evaluation functions I planned to applyto my Use Case model:

* Apply a K-Fold Cross Validation (or possibly a StratifiedKFold) function.
* Calculate a confusion matrix, analyzing precision, recall and F1.
* Calculate a ROC curve.

Only to then realize, these all apply to *classifiers*, and not *regressors.* Arg… I’ve since amended my approach to apply these…

* Calculate the Mean Squared Error
* Calculate the Coefficient of Determination (R squared)

I should probably take a step back on Monday and review both types of models as well as their evaluation techniques. I realize that, as a Data Scientist, I’ll need to keep all these straight, so this has been a good learning experience.