11/9 pm

Model1.ipynb through model5 preliminary runs for comparison.

**model0\_compare.ipynb** – runs all features through nine different models each with no scaling, standard scaling, and minmax scaling. After each step a row is appended to a data frame displaying model, scaling and score (training data, testing data, and an average). All models ran relatively quick except for the svc model with no scaling, which is why that block of code is commented out.

Feature selection needs to be done on the top performing models.

Probably with rfe…

<https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html>

<https://machinelearningmastery.com/rfe-feature-selection-in-python/>

or

<https://towardsdatascience.com/feature-selection-techniques-in-machine-learning-with-python-f24e7da3f36e>

<https://towardsdatascience.com/explain-your-machine-learning-with-feature-importance-774cd72abe>

<https://machinelearningmastery.com/calculate-feature-importance-with-python/>

<https://towardsdatascience.com/backward-elimination-for-feature-selection-in-machine-learning-c6a3a8f8cef4>

Classifier parameters need to be gridsearched and modified for best reults.

11/10 pm

Go thru code one more

Write standalone to read in saved file, rerun split, fit

Run with candidates to predict… also put in module0 for three final models

Create output xl or csv with predictions

Work on readme doc

Go thru specs and rubric

All models in one pgm…?

Download csv from nasa and run against.

11/12

Running 3 models against 1687 candidate records

DT: 1665 confirmed, 22 false positive

RF: 1515 confirmed, 172 false positive

SVC: 3 confirmed, 1684 false positive

Choosing random forest as it is the least biased.