# Process for building a data model

1) Query, ingest, load - (this is stored in jupyter notebook 1)

2) Preprocessing - (this is stored in jupyter notebook 2)

3) Train Test Split + Modeling –

Sample dataset down to 100 or so data points

Preprocessing is a class that can store key features, such as mean, std, etc

Once you’ve demonstrated business value, switch Jupyter notebook with preprocessing into an importable package

# General jupyter advice

For each notebook, start with goals summary at the top.

Bullet 2

Pip installable (“extensions”)

Table of contents add in

Scratchpad add in

# General Atom advice

# General Python advice

Python has something like R Studio that was just started called “Jupyter lab’

Share packages you develop with team – don’t need to reinvent the wheel

Use Numpy vectors for lin alg only - pass data frames to the actual function

# General modeling advice

Confusion matrices

Box example - 1,0 0,1 -- most often care about the ones you get WRONG -

# Some next steps for this model

Start with Logistic Regression

However, expect that a random forest will predict better, because of non-linear behavior and separation boundaries

Random forest – will be used to understand the data set and understand feature importance

Can use it to answer: do I need to take these measurements out of the dataset?

# Next time

~~Fix problem with importing packages~~

Install library to view Jupyter

AUTORELOAD - look into this

PDB - look into this

Use PDB with scratchpad

My goals:

implement ridge regression

Categorical to continuous regression...?

Demonstrate that I can replicate code in SKlearn

Try to build a cool UI out of these...