Feature Engineering

Chapter 2

Stephen Kimel

Starting point:

- Outcome measurement Y (also called dependent variable, response, target).
- Vector of p predictor measurements X (also called inputs, regressors, covariates, features, independent variables).
- In the *regression problem*, Y is quantitative (e.g price, blood pressure).
- In the *classification problem*, Y takes values in a finite, unordered set (survived/died, digit 0-9, cancer class of tissue sample).
- We have training data $(x_1, y_1), \ldots, (x_N, y_N)$. These are observations (examples, instances) of these measurements.

Source: ISLR http://faculty.marshall.usc.edu/gareth-james/ISL/

Important concepts

Which of the below models does a better job in this case?

Model Bias and Variance

- **Variance:** How much would *f* would change if we estimated it with a different training dataset. (Associated with overfitting.)
- **Bias:** Error introduced by approximating a real-life problem (complicated) by a much simpler model. (Associated with underfitting)

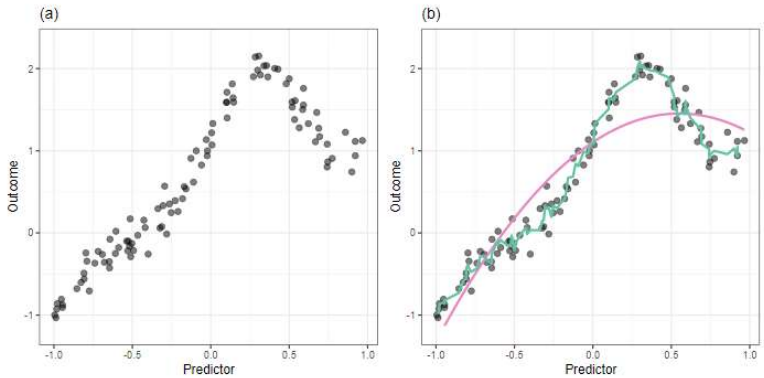


Figure 1.5: A simulated data set and model fits for a 3-point moving average (green) and quadratic regression (purple).

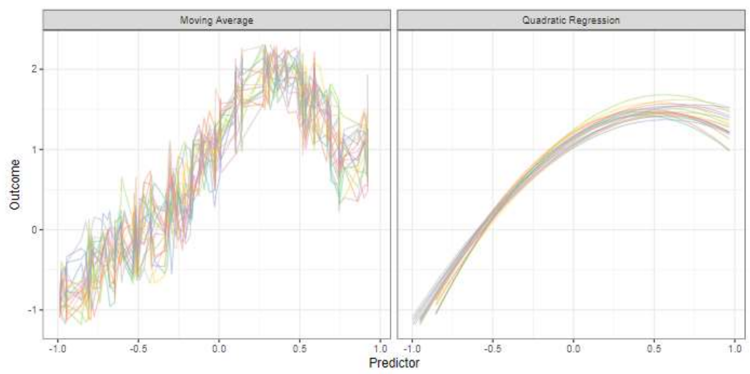


Figure 1.6: Model fits for twenty jittered versions of the data set.

Logistic Regression

Logistic Regression Attributes

- Used for classification (not regression!)
- High Bias & Low Variance
- Interpretable

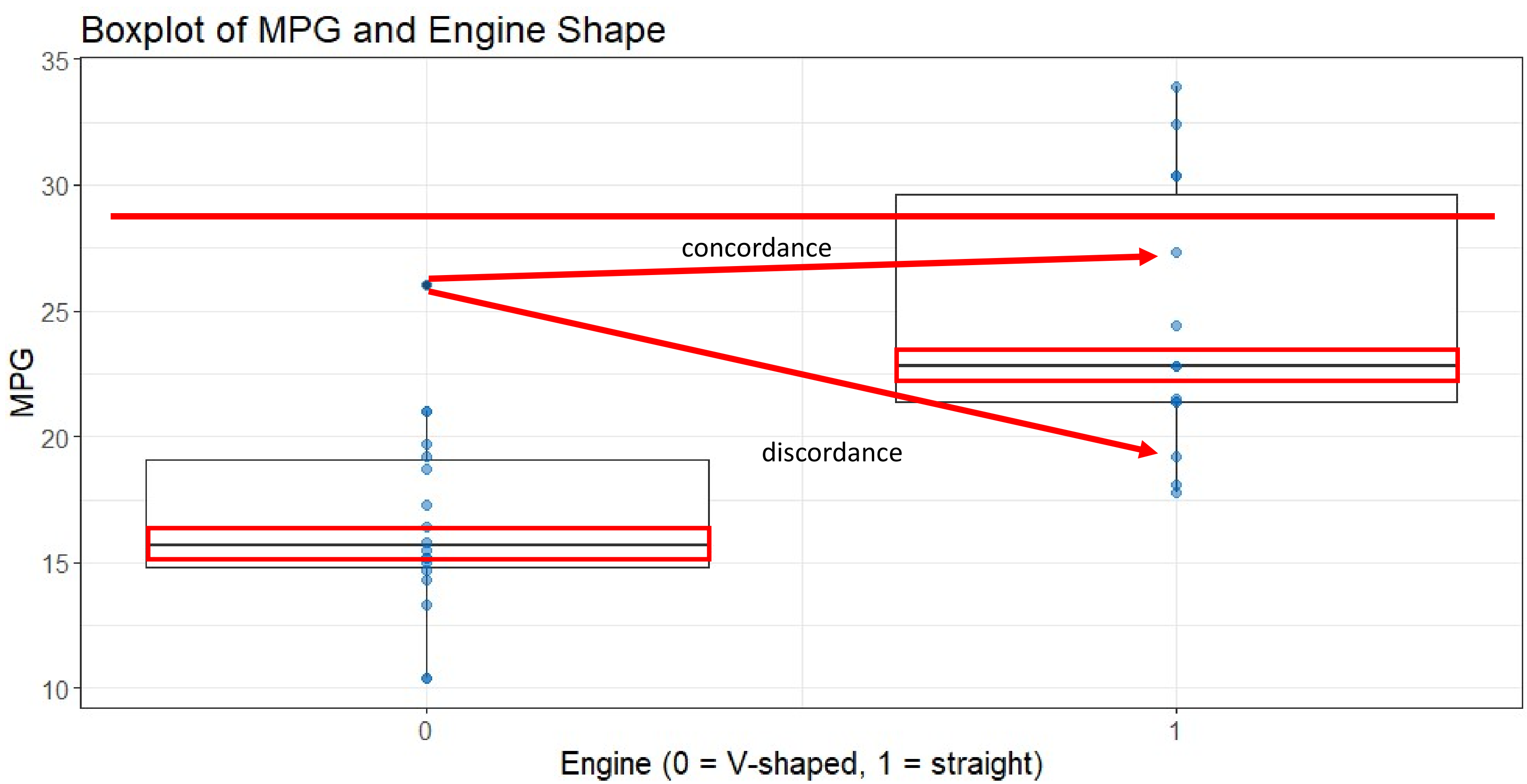
Interpretable

Simple Equation

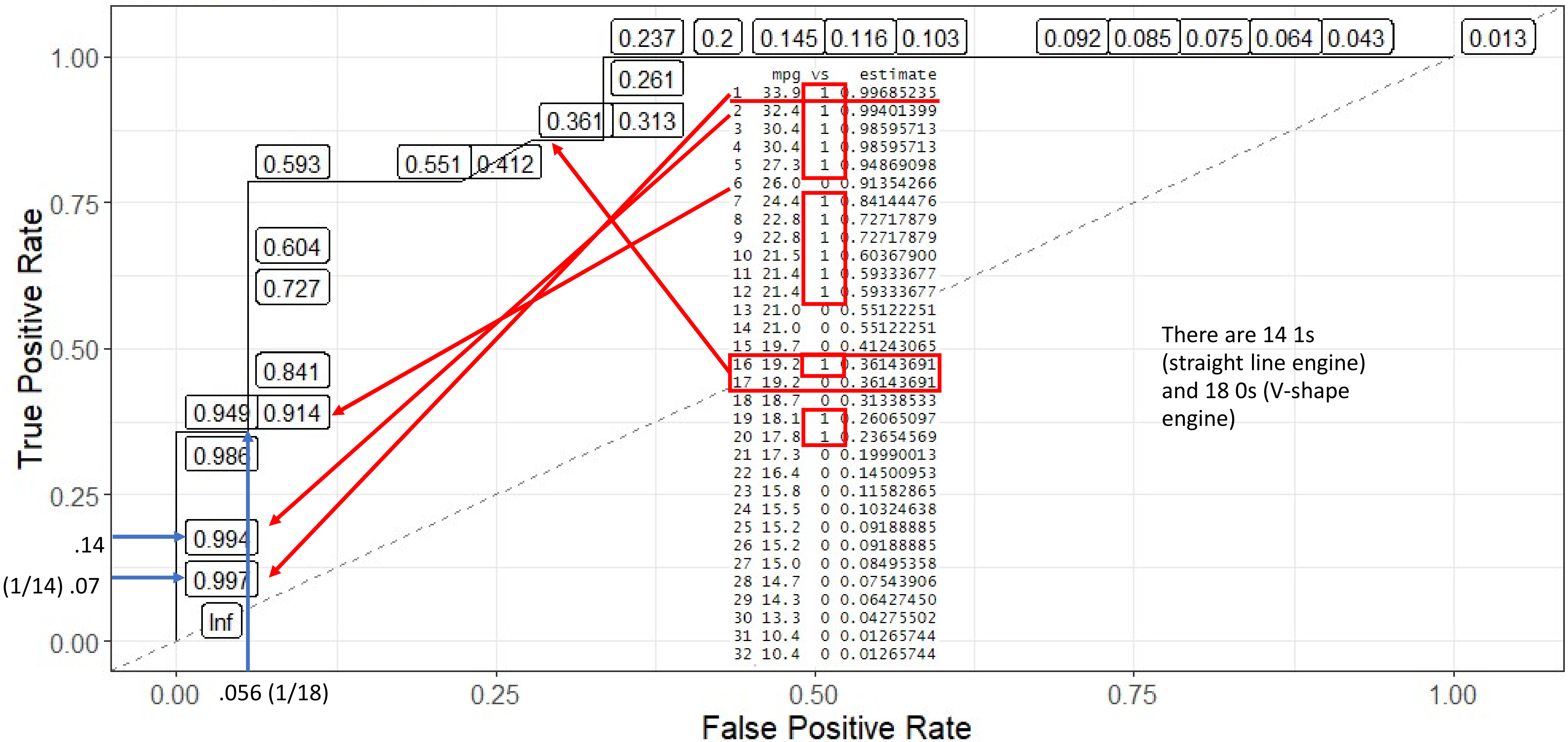
- $y = -8.8 + .43 \text{mpg}$

Thing we are interested in predicting. For logistic regression, this will measure how likely an observation is in a certain bucket of interest.

	mpg	vs
Mazda RX4	21.0	0
Mazda RX4 Wag	21.0	0
Datsun 710	22.8	1
Hornet 4 Drive	21.4	1
Hornet Sportabout	18.7	0
Valiant	18.1	1
Duster 360	14.3	0
Merc 240D	24.4	1
Merc 230	22.8	1
Merc 280	19.2	1
Merc 280C	17.8	1
Merc 450SE	16.4	0
Merc 450SL	17.3	0
Merc 450SLC	15.2	0
Cadillac Fleetwood	10.4	0
Lincoln Continental	10.4	0
Chrysler Imperial	14.7	0
Fiat 128	32.4	1
Honda Civic	30.4	1
Toyota Corolla	33.9	1
Toyota Corona	21.5	1
Dodge Challenger	15.5	0
AMC Javelin	15.2	0
Camaro Z28	13.3	0
Pontiac Firebird	19.2	0
Fiat X1-9	27.3	1
Porsche 914-2	26.0	0
Lotus Europa	30.4	1
Ford Pantera L	15.8	0
Ferrari Dino	19.7	0
Maserati Bora	15.0	0
Volvo 142E	21.4	1



AUC for Engine Size



	mpg	vs	estimate
1	33.9	1	0.99685235
2	32.4	1	0.99401399
3	30.4	1	0.98595713
4	30.4	1	0.98595713
5	27.3	1	0.94869098
6	26.0	0	0.91354266
7	24.4	1	0.84144476
8	22.8	1	0.72717879
9	22.8	1	0.72717879
10	21.5	1	0.60367900
11	21.4	1	0.59333677
12	21.4	1	0.59333677
13	21.0	0	0.55122251
14	21.0	0	0.55122251
15	19.7	0	0.41243065
16	19.2	1	0.36143691
17	19.2	0	0.36143691
18	18.7	0	0.31338533
19	18.1	1	0.26065097
20	17.8	1	0.23654569
21	17.3	0	0.19990013
22	16.4	0	0.14500953
23	15.8	0	0.11582865
24	15.5	0	0.10324638
25	15.2	0	0.09188885
26	15.2	0	0.09188885
27	15.0	0	0.08495358
28	14.7	0	0.07543906
29	14.3	0	0.06427450
30	13.3	0	0.04275502
31	10.4	0	0.01265744
32	10.4	0	0.01265744

There are 14 1s (straight line engine) and 18 0s (V-shape engine)

Chapter 2

Example in Book

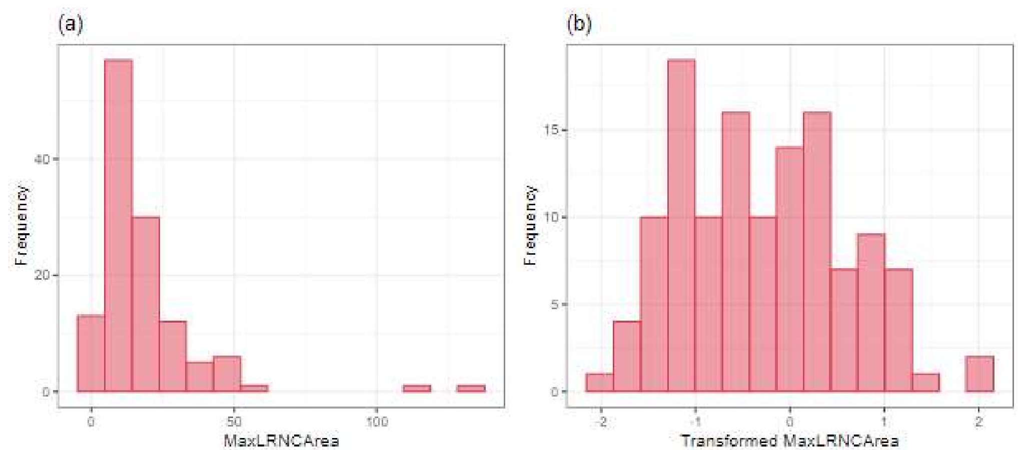
- Predict patient risk for ischemic stroke
- Historically just used size of blockage to predict
  - Shown to be a poor predictor
- Perhaps the type of blockage will help the model
- Splitting data

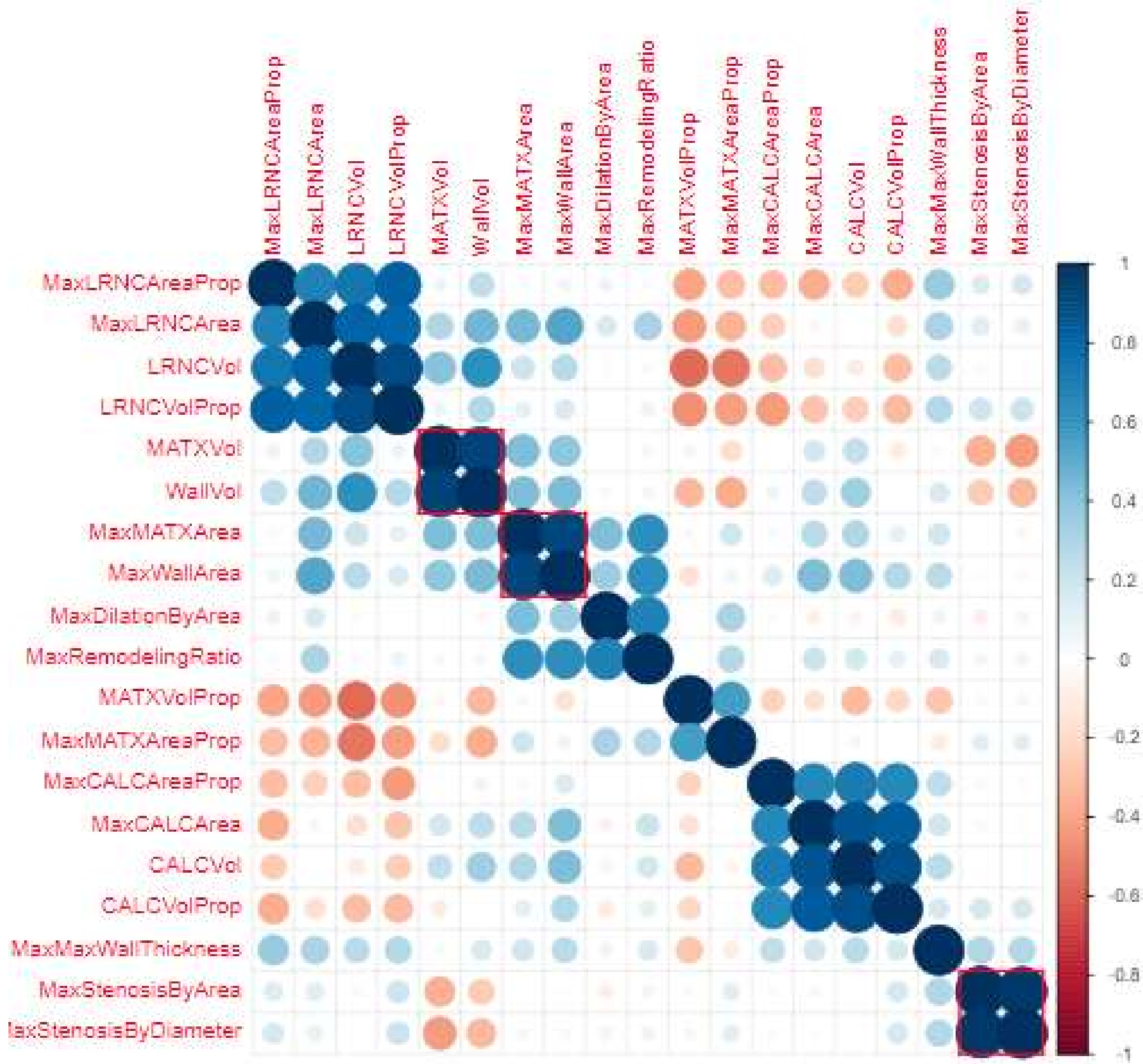
Table 2.2: Distribution of stroke outcome by training and test split.

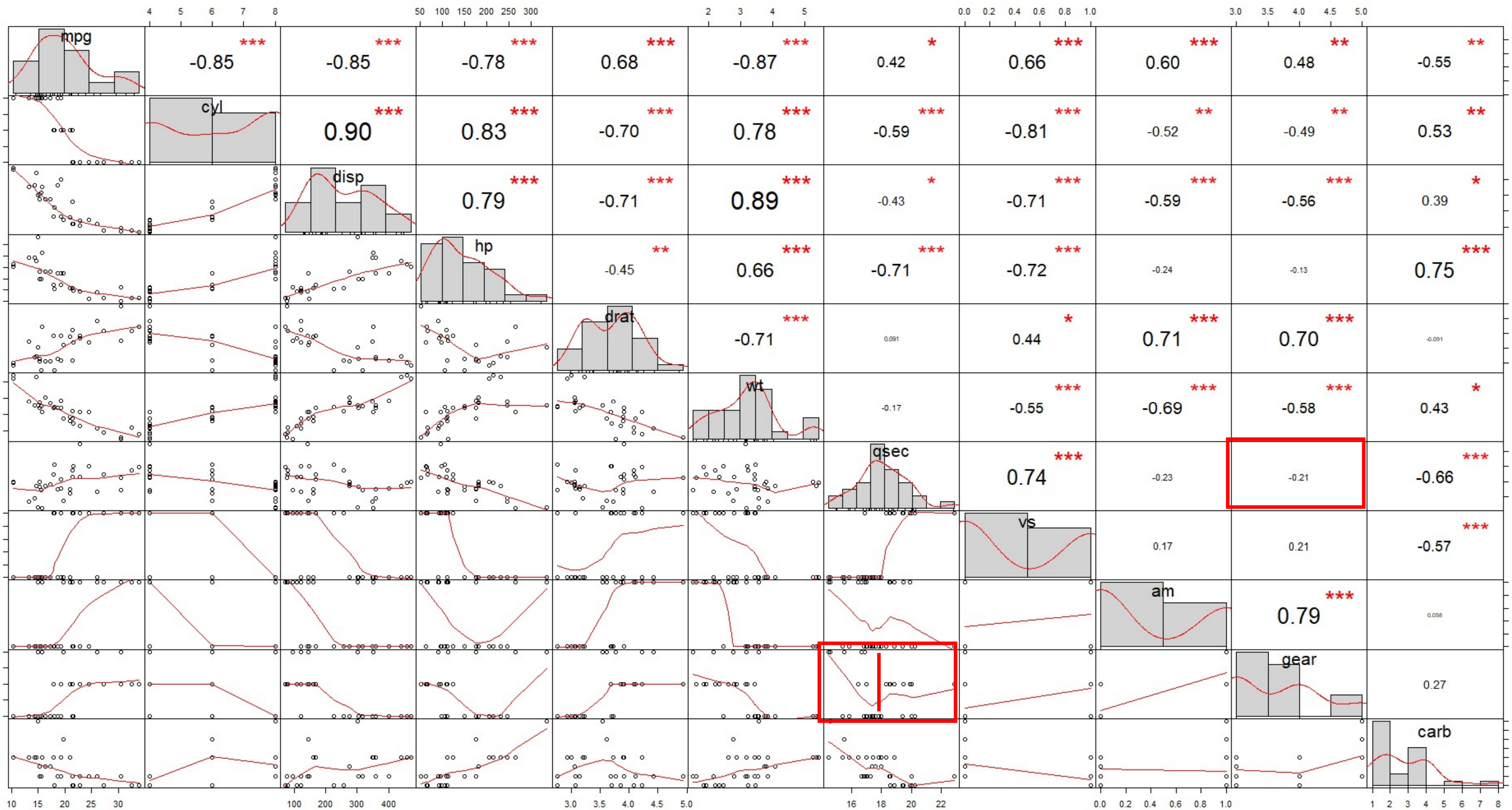
Data Set	Stroke = Yes (n)	Stroke = No (n)
Train	51% (45)	49% (44)
Test	51% (19)	49% (18)

Preprocessing Data

- Input distributions
- Missing data (Chapter 8) – some models can handle missing data, some can't
- Unusual values (Chapter 6)
  - Should you remove them, transform them, or leave them?
- Relationships between inputs
- Relationship of inputs with response




MaxLRNCAreaProp
MaxLRNCArea
LRNCVol
LRNCVolProp
MATXVol
WallVol
MaxMATXArea
MaxWallArea
MaxDilationByArea
MaxRemodelingRatio
MATXVolProp
MaxMATXAreaProp
MaxCALCAreaProp
MaxCALCArea
CALCVol
CALCVolProp
MaxMaxWallThickness
MaxStenosisByArea
MaxStenosisByDiameter
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1



mpg
cyl
disp
hp
drat
wt
qsec
vs
am
gear
carb
-0.85 ***
-0.85 ***
-0.78 ***
0.68 ***
-0.87 ***
0.42 *
0.66 ***
0.60 ***
0.48 **
-0.55 **
0.90 ***
0.83 ***
-0.70 ***
0.78 ***
-0.59 ***
-0.81 ***
-0.52 **
-0.49 **
0.53 **
0.79 ***
-0.71 ***
0.89 ***
-0.43 *
-0.71 ***
-0.59 ***
-0.56 ***
0.39 *
-0.45 **
0.66 ***
-0.71 ***
-0.72 ***
-0.24
-0.13
0.75 ***
-0.71 ***
0.091
0.44 *
0.71 ***
0.70 ***
-0.091
-0.17
-0.55 ***
-0.69 ***
-0.58 ***
0.43 *
0.74 ***
-0.23
-0.21
-0.66 ***
0.17
0.21
-0.57 ***
0.79 ***
0.058
0.27

Exploration

for *each resample* **do**

  Use the resample's 90% to fit models M_1 and M_2

  Predict the remaining 10% for both models

  Compute the area under the ROC curve for M_1 and M_2

  Determine the difference in the two AUC values

end

Use a one sided t-test on the differences to test that M_2 is better than M_1.

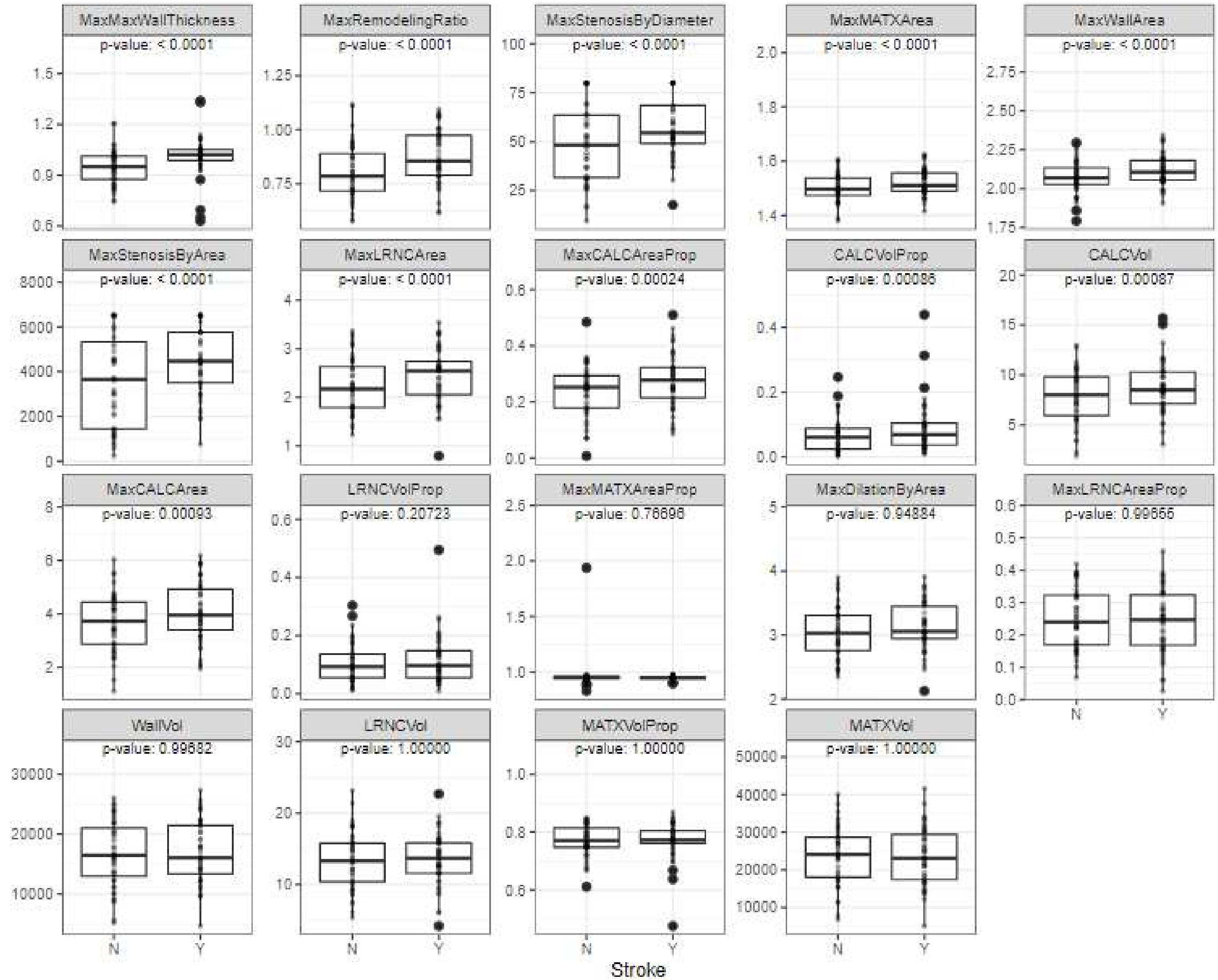
Should the procedure above happen before splitting the data or afterwards?

Divide data into K roughly equal-sized parts ($K = 5$ here)

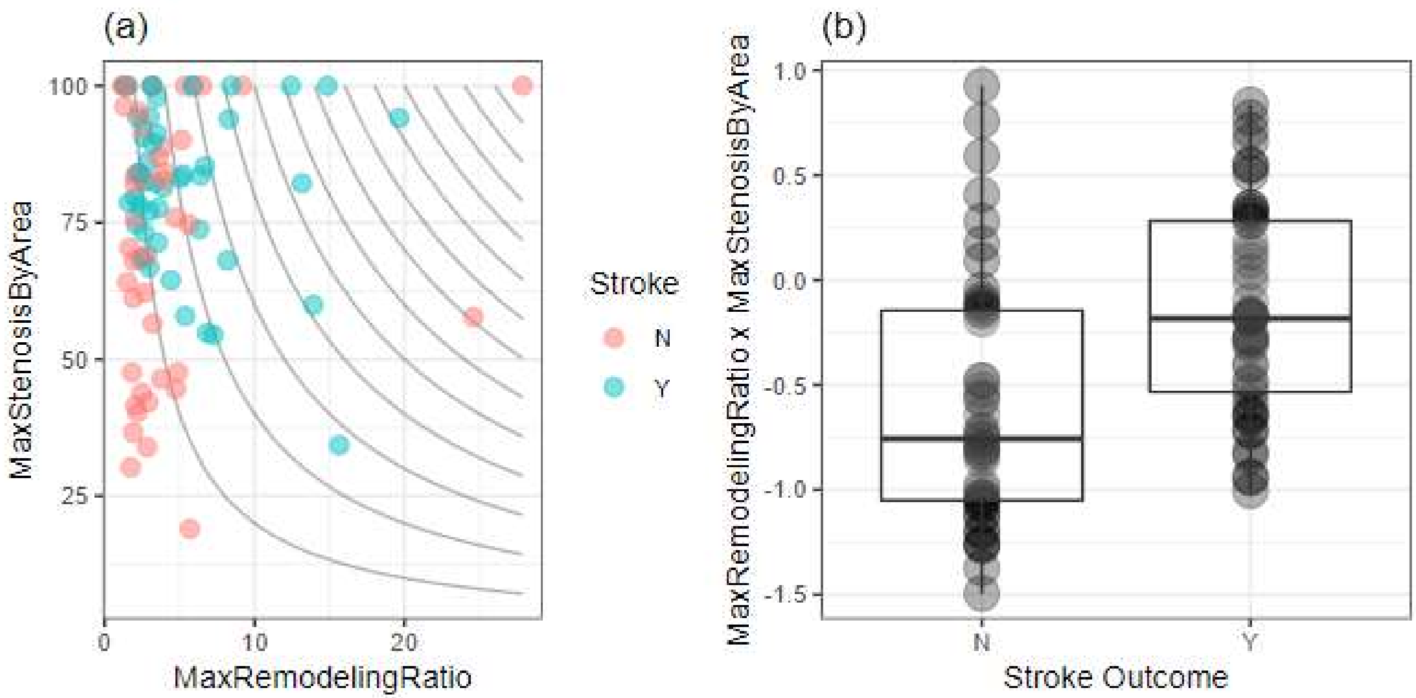
1	2	3	4	5
Validation	Train	Train	Train	Train
Train	Validation	Train	Train	Train

…

p-value = the probability that our assumption is true given the data we observed



Interactions



Choosing Input Variables

- Chapters 10 & 11
- “Using this training set, we estimated that the filtered predictor set of 7 imaging predictors was our best bet.”
- “How well did this predictor set do on the test set? The test set area under the ROC curve was estimated to be 0.69. This is less than the resampled estimate of 0.72 but is greater than the estimated 90% lower bound on this number (0.674).”

Is it a good idea to back and build more models and retest? Why?