Primer on Machine Learning

attilagk

Lazy8

Contents

Machine Learning is Everywhere

- pattern detection and recognition (iphone touch ID, face ID, word autocomplete, speech to text)
- history based recommendation (youtube, facebook, google search, amazon,...)
  - products for customers
  - customers for providers
- email filtering and classification (gmail)

Machine Learning and Artificial Intelligence

1950s *[getting] machines to exhibit behavior, which if done by humans, would be assumed to involve the use of intelligence*[1]

now *computational methods to automatically learn and to improve with experience*[2]

ML[3] statistical ("statistical learning")

AI analytical (knowledge, logic)

[1] Arthur Samuel, 1983

[2] http://www.mlplatform.nl/what-is-machine-learning/

[3] may mean Maximum Likelihood: abbrev. not widely used

Machine Learning Now

- big data
  - data science, data mining, ...
  - myth: machine learning needs big data[4]
- fast computers
- emerging new methods
  - deep learning, reinfocement learning, ...

[4] see NY/SF homes toy data

General Outline

1. formulate task (EN, ST)
2. collect data (EN, ST)
3. "engineer" data (EN, PR, DSS, HPC)
4. modeling (ST, EN)
5. fit/learn/train model(s) on data (LIB, PR, HPC)
6. validate, select (LIB, ST, PR, HPC)
7. apply model to test data (LIB)
8. interpret results (EN, ST)
9. report (DOC)

skill set	
EN	expert knowledge
ST	statistics
PR	programming
DSS	domain spec. software
HPC	high performance computing
LIB	ML libraries[a]
DOC	LaTeX, Web

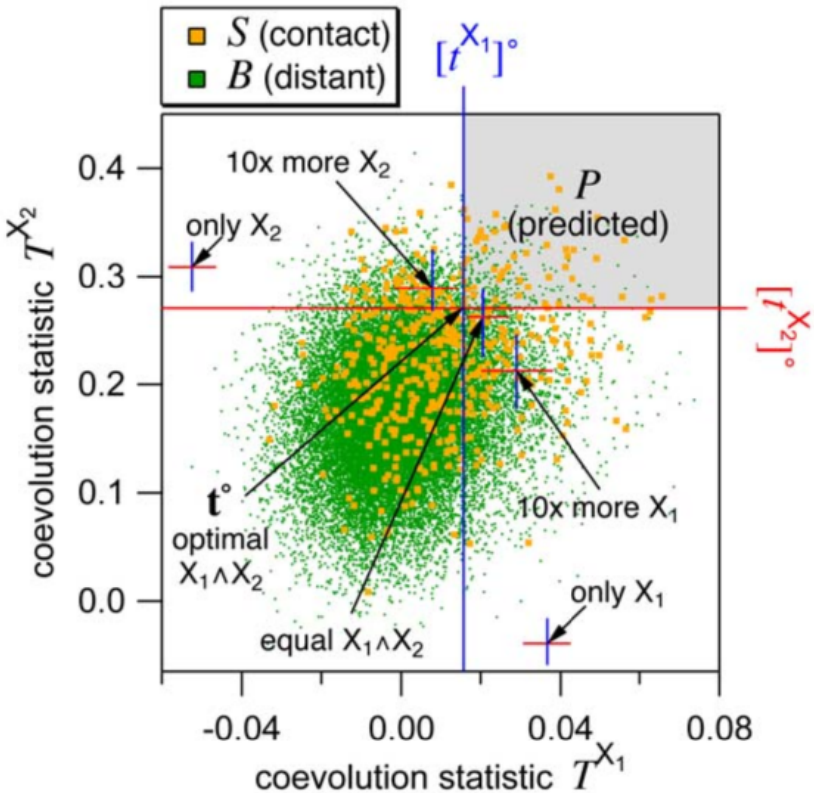
[a] Python, R, Java, Julia, Scala, Matlab

My Story with Machine Learning

Myth: it's like cooking

skill	2006	2017
expert knowledge	?	?
statistics	-	+
programming	-	+
domain spec. softw.	-	+
high perform. comp.	-	?
ML libraries	-	?
LaTeX, Web.	-	+

PLoS One. 2012;7(5):e36546.



Contents

Tasks

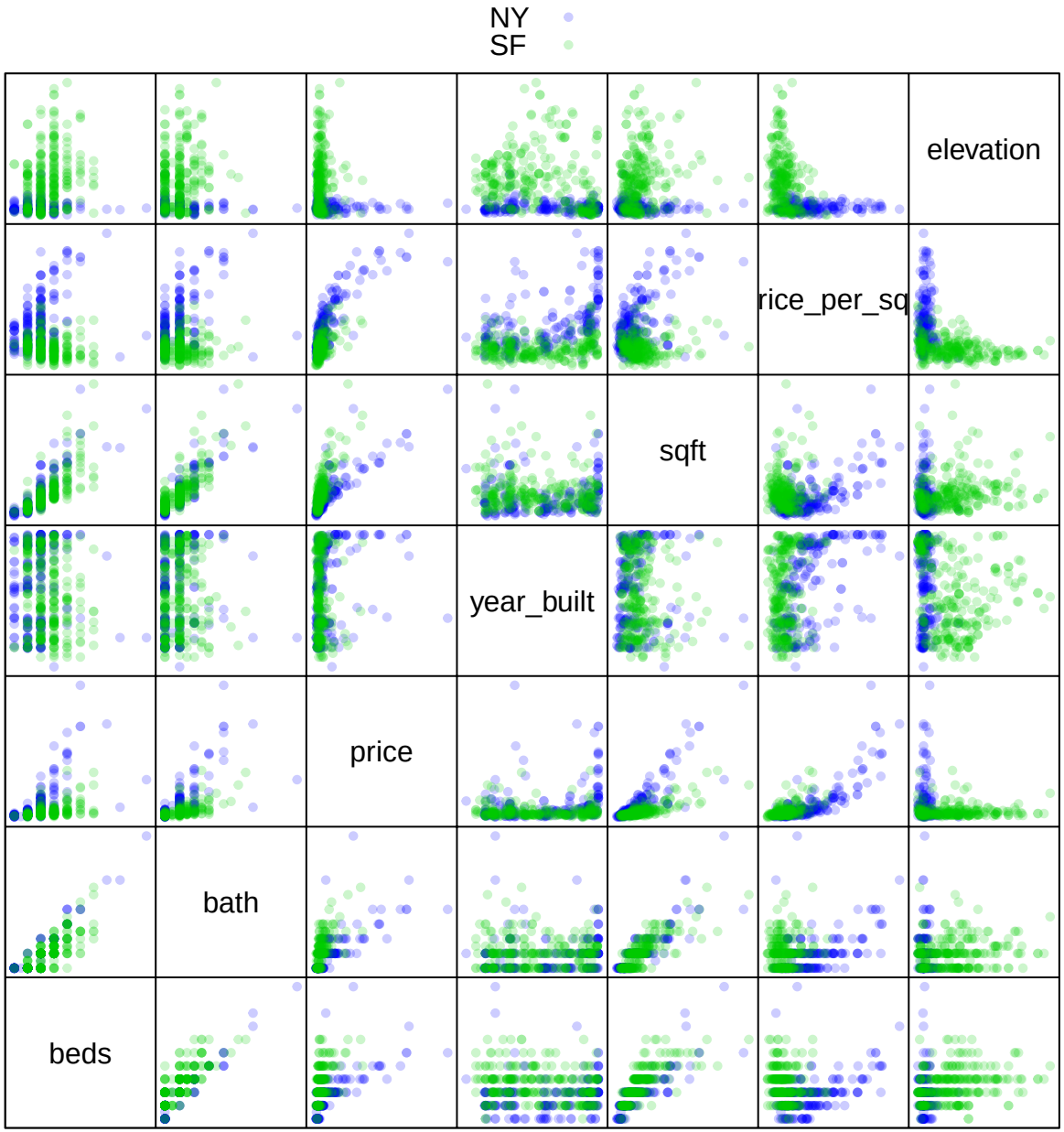
1. supervised learning: training *and* test data
   - prediction, classification
     - pattern recognition
     - business, medical, ... predictions & decisions
2. unsupervised learning: *only* training data
   - structure discovery
     - social, biol., tech. networks, associations,...
     - probabilistic expert systems
   - hypothesis testing, feature subset selection
     - research, marketing
   - matrix completion (imputation)
     - recommendation systems

The "Home" Data for Classification

Useless except for demonstration

observation	input features / variables				output
i	x_{i1}	x_{i2}	...	x_{ip}	y_i
home	price/sqft	elevation	...	beds	city
training data					
1	999	10	...	2	NY
2	1939	0	...	2	NY
⋮	⋮	⋮	⋮	⋮	⋮
491	764	163	...	1	SF
492	762	216	...	3	SF
test data					
493	1196	40	...	2	?
⋮	⋮	⋮	⋮	⋮	⋮

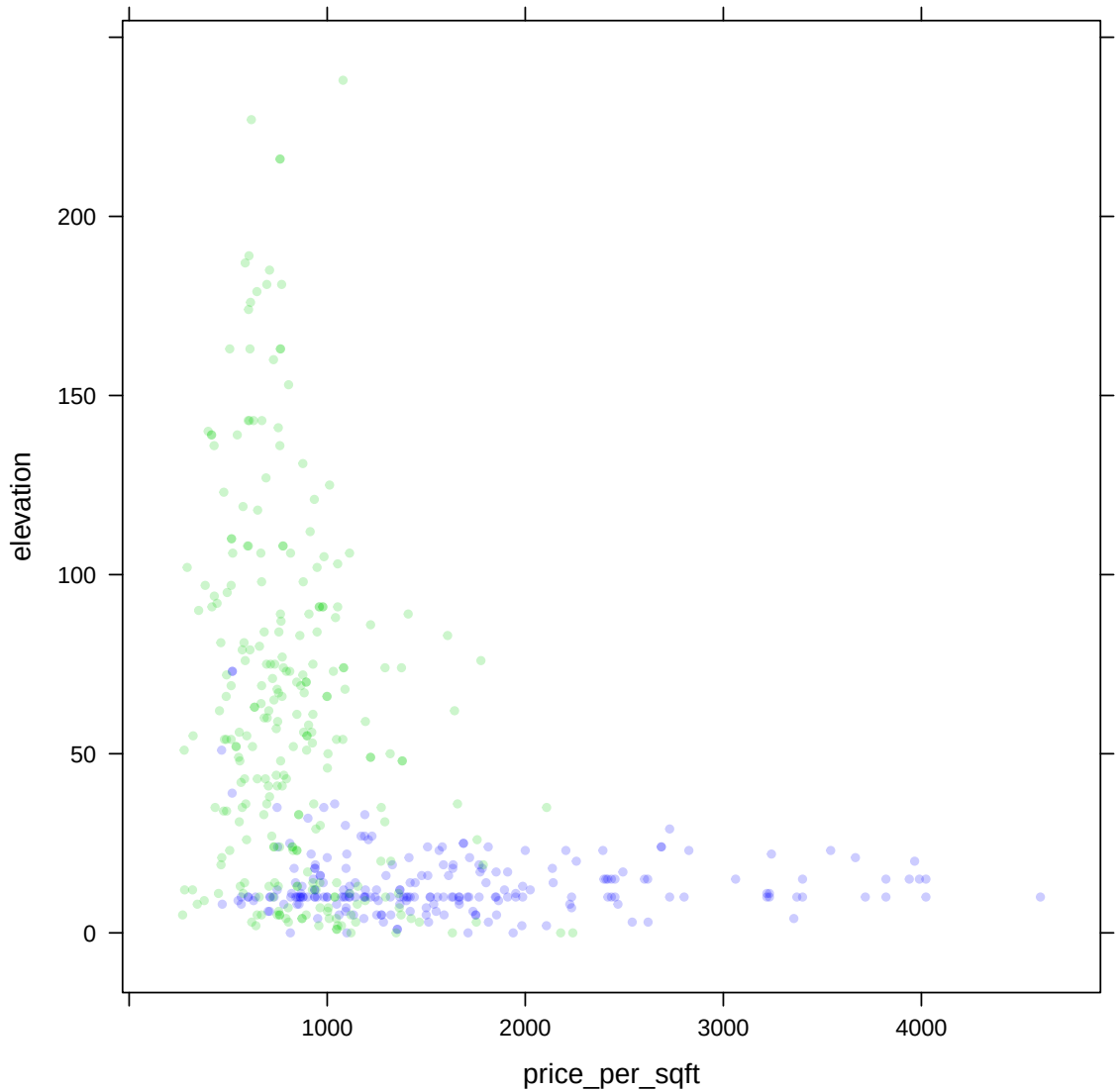
Home Data Overview



Scatter Plot Matrix

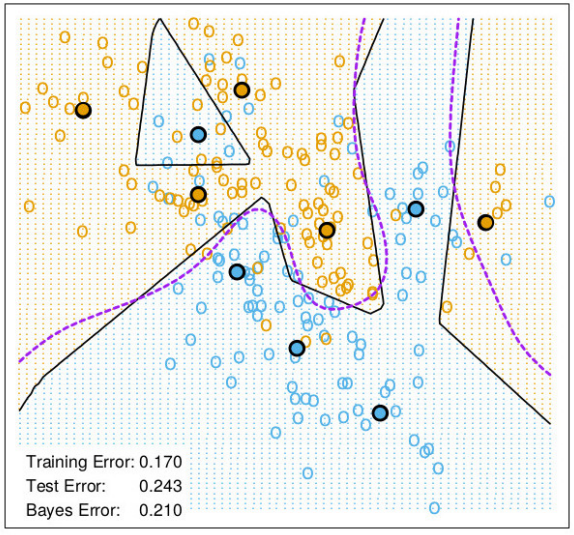
Informative Features

2 input features: 2D plots

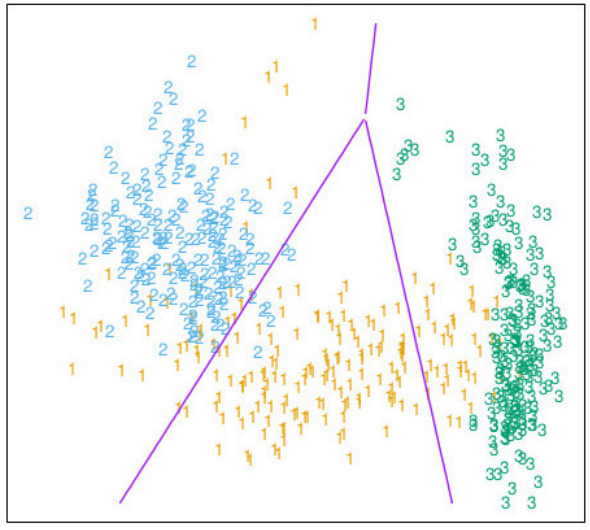


Models for Classification

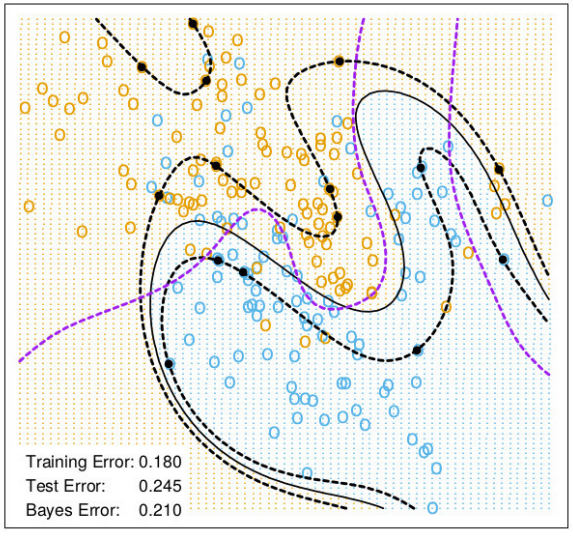
K-means classifier



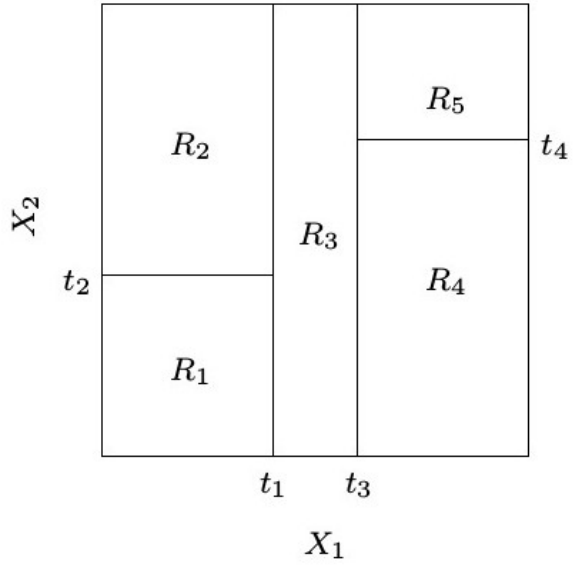
generalized linear regression



support vector machine



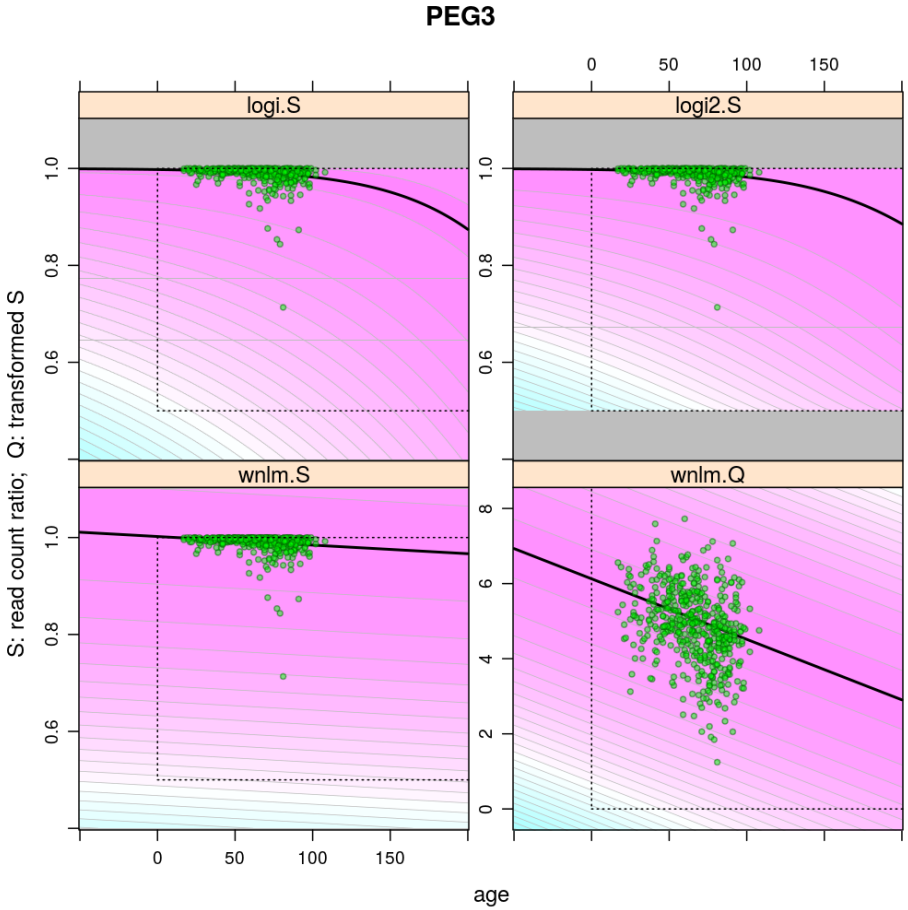
decision tree (poor perform.)



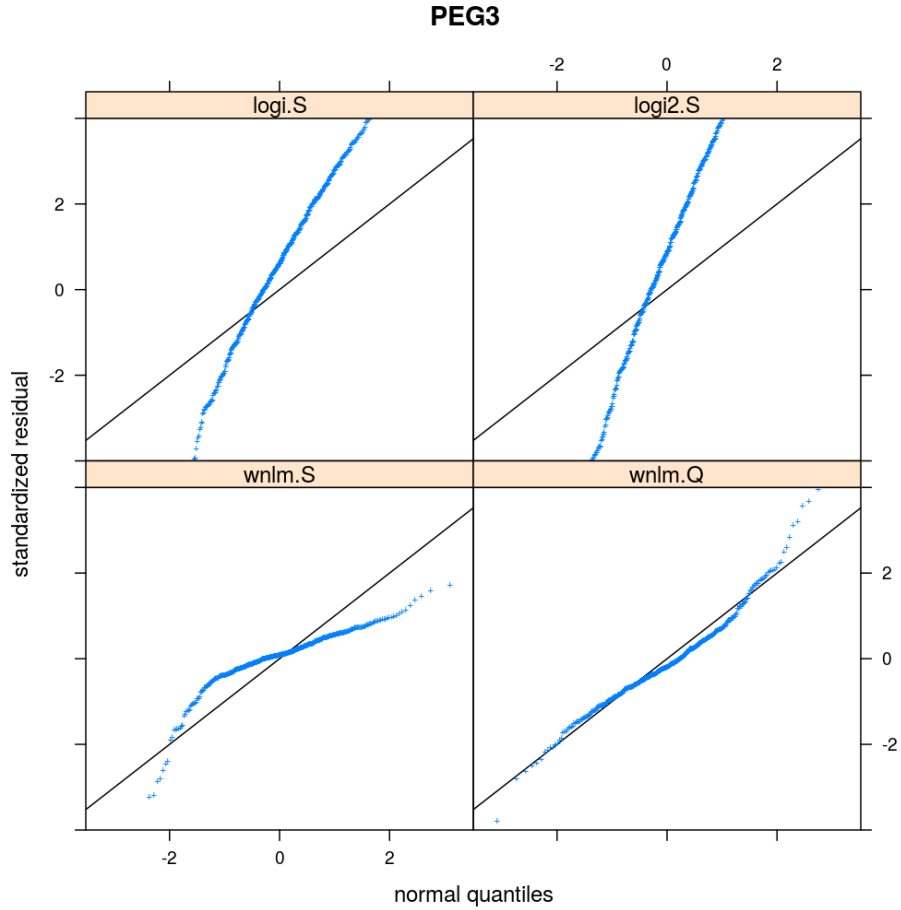
Which Model to Select? The One That Fits the Best![5]

Information theory, optimality, model ensemble

subjective evaluation



diagnostics



[5] should also be: simple, interpretable, fast to fit,...

Contents

Building on "Visual Intro..."[6]

what we will address in addition to "Visual Intro..."

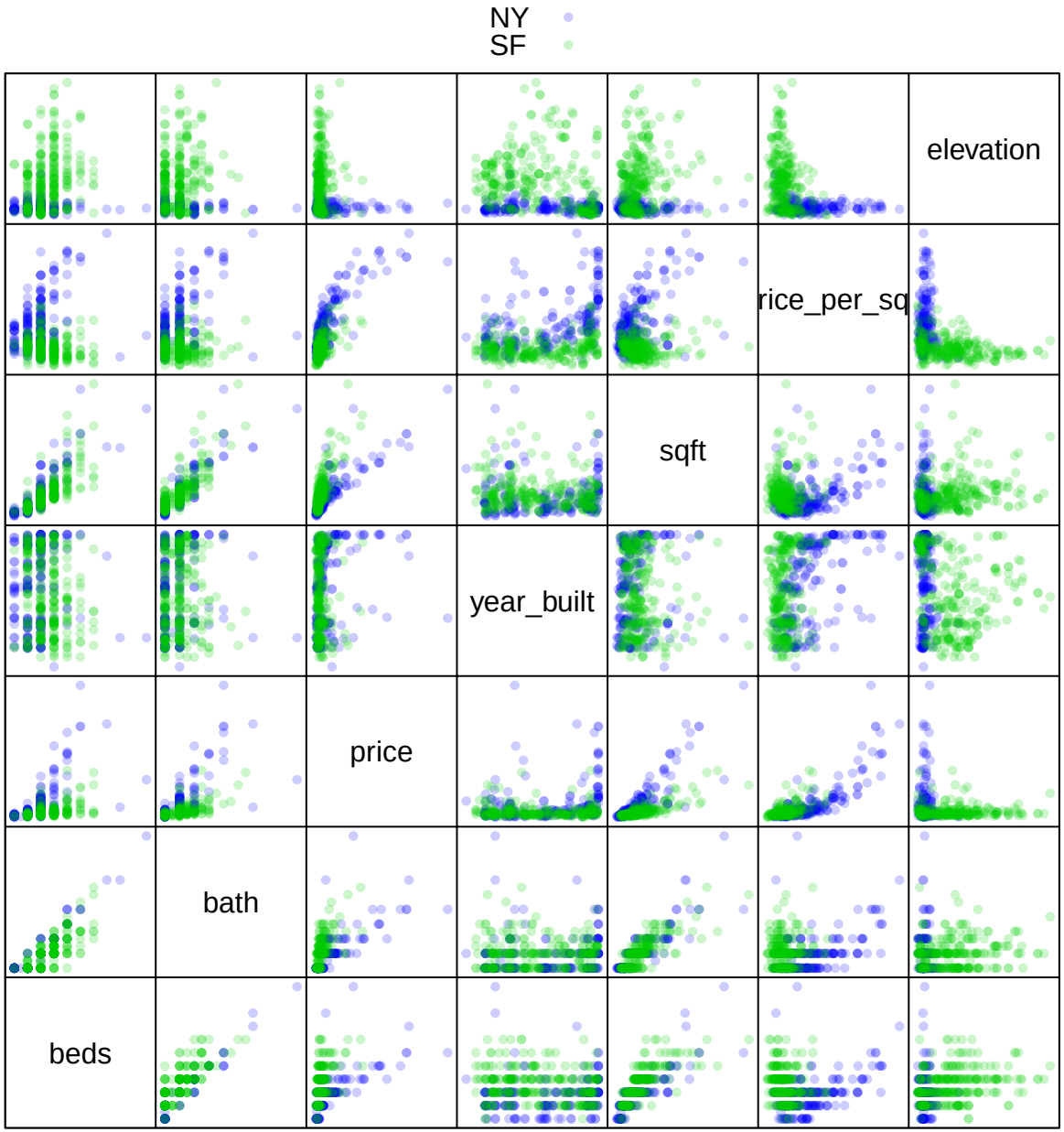
- model validation/selection
- prediction

what we won't

- real World task
- data collection/engineering
- model selection
- validation, performance

[6] http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

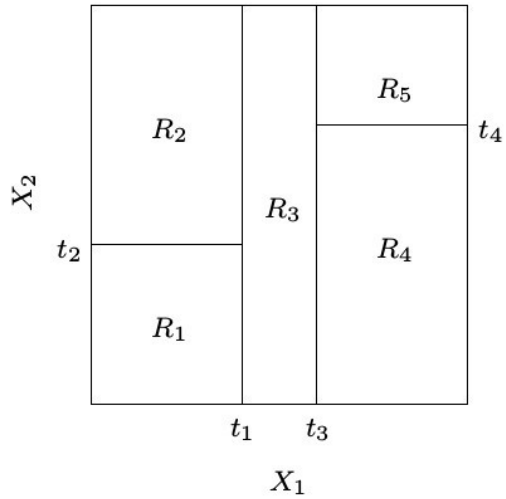
Home Data Overview

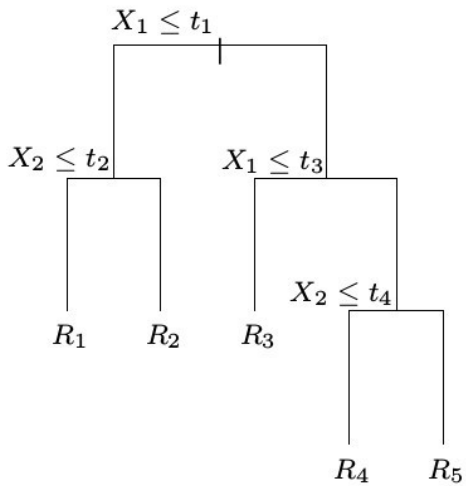


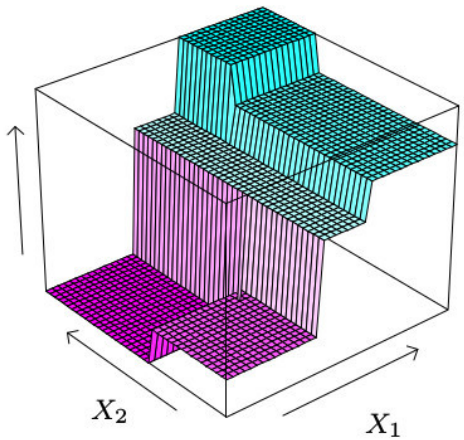
Scatter Plot Matrix

Decision Tree is Ideal for Demo

A.k.a. CART: Classification And Regression Tree







Demo with "Visual Intro"[7]

Observe progressive growth of tree!

[7] http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

Our Analysis; Why R?

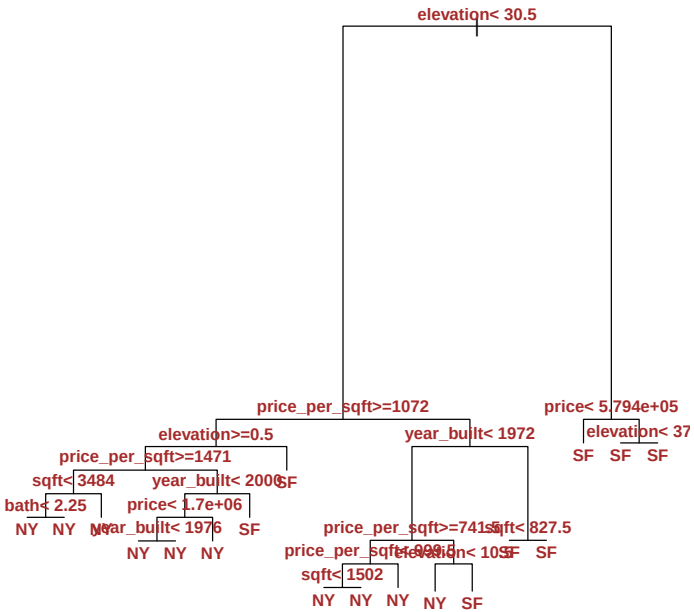
- created by and for biostatisticians
- functional language (like JavaScript)
- open source
- mature
- lots of machine learning packages
- R2D3[8]

[8] http://www.r2d3.us/visual-intro-to-machine-learning-part-1/

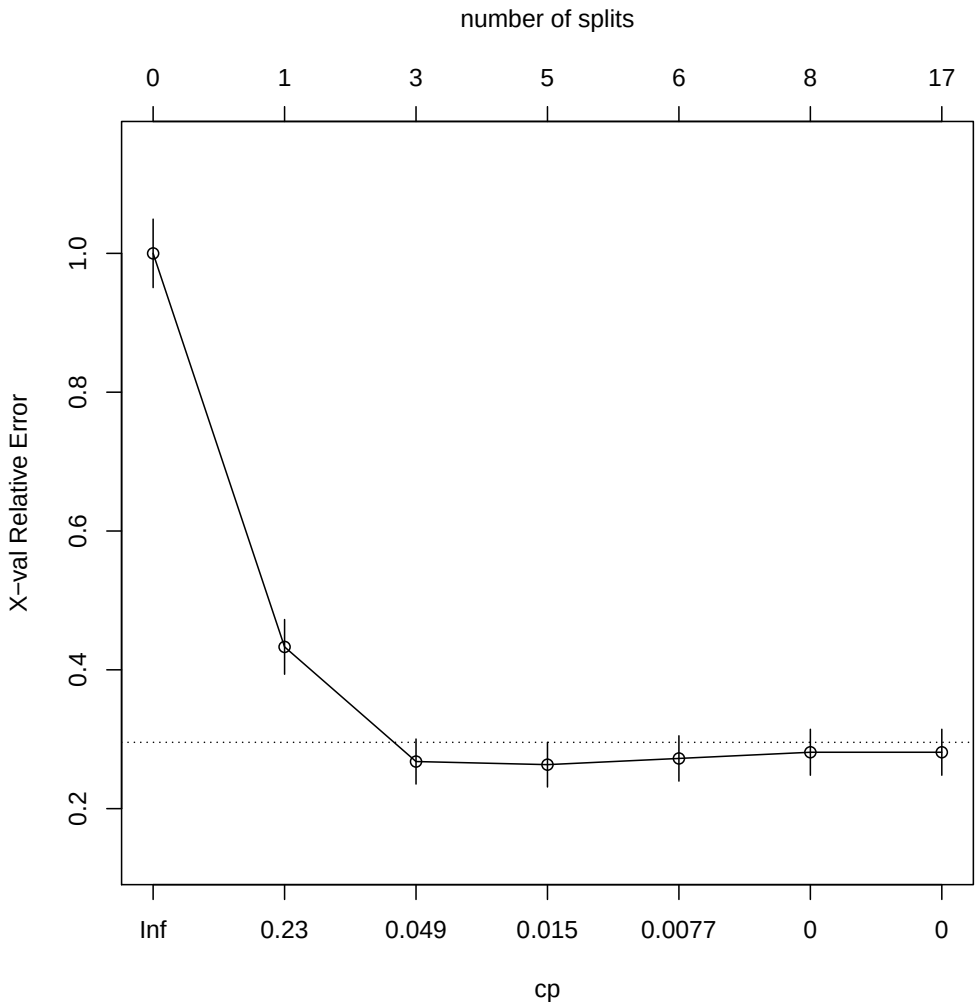
Fitting Decision Trees

Using the `rpart` R library

a complex tree



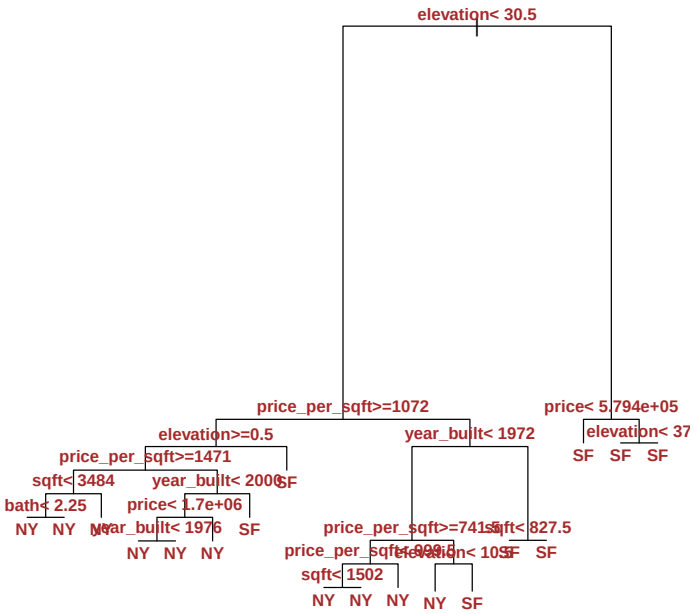
Avoiding Overfitting (Too Much Complexity)



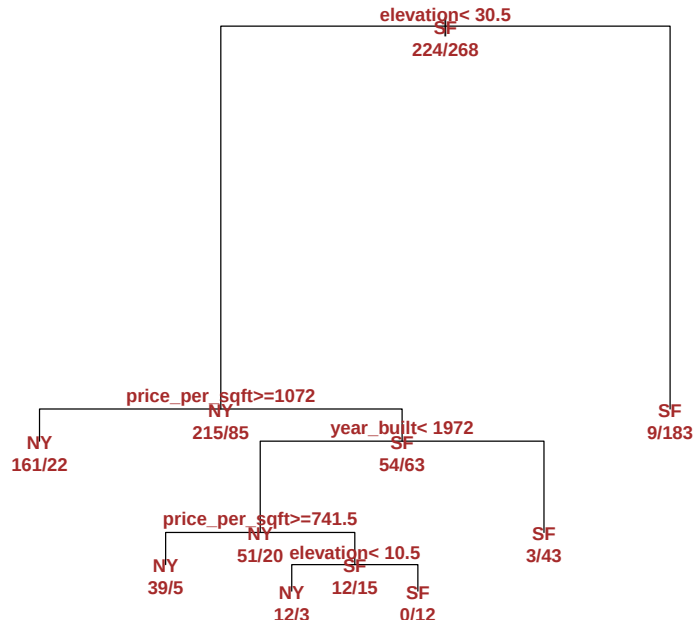
Fitting Decision Trees

Using the `rpart` R library

a complex tree

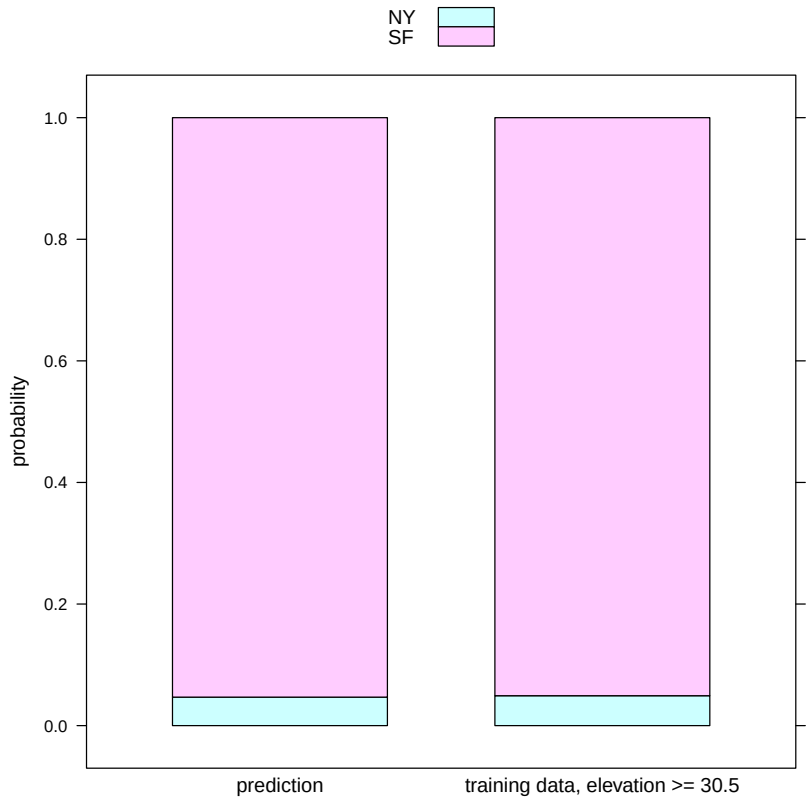


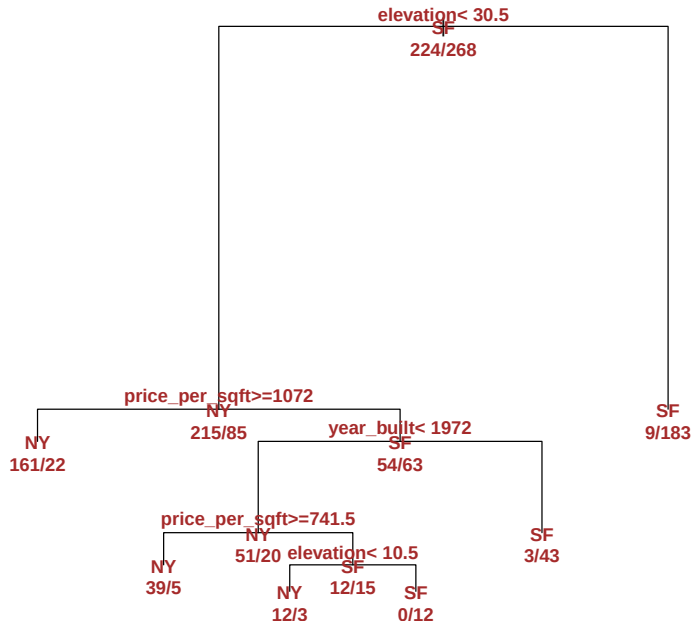
the optimal tree



Prediction

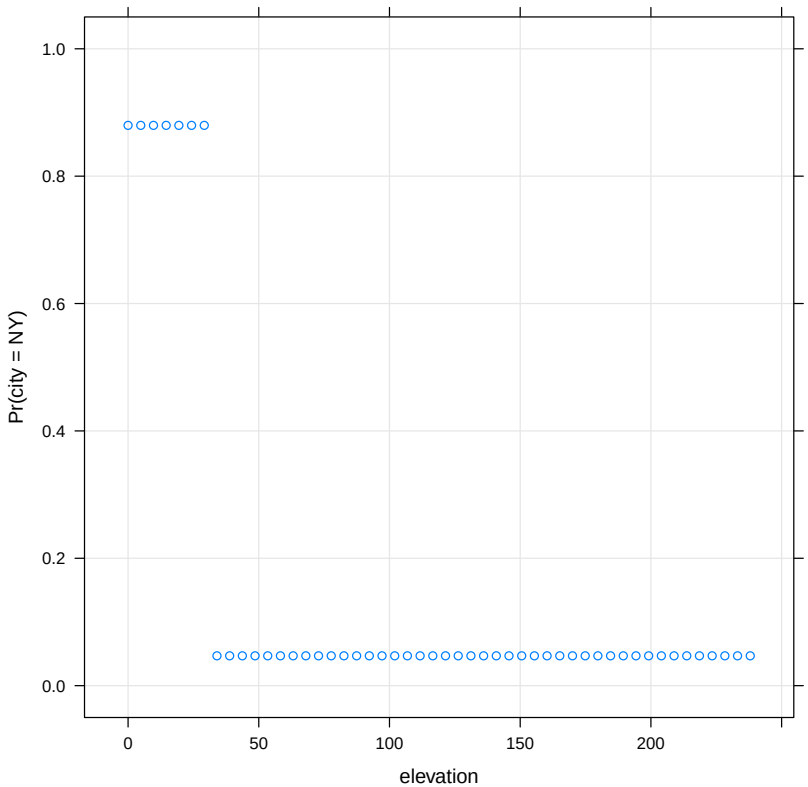
at “the average of traning data”

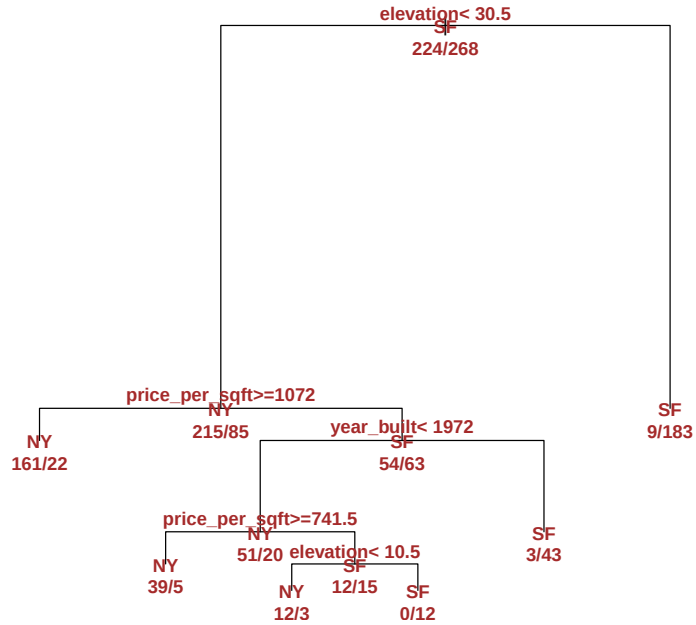




Prediction

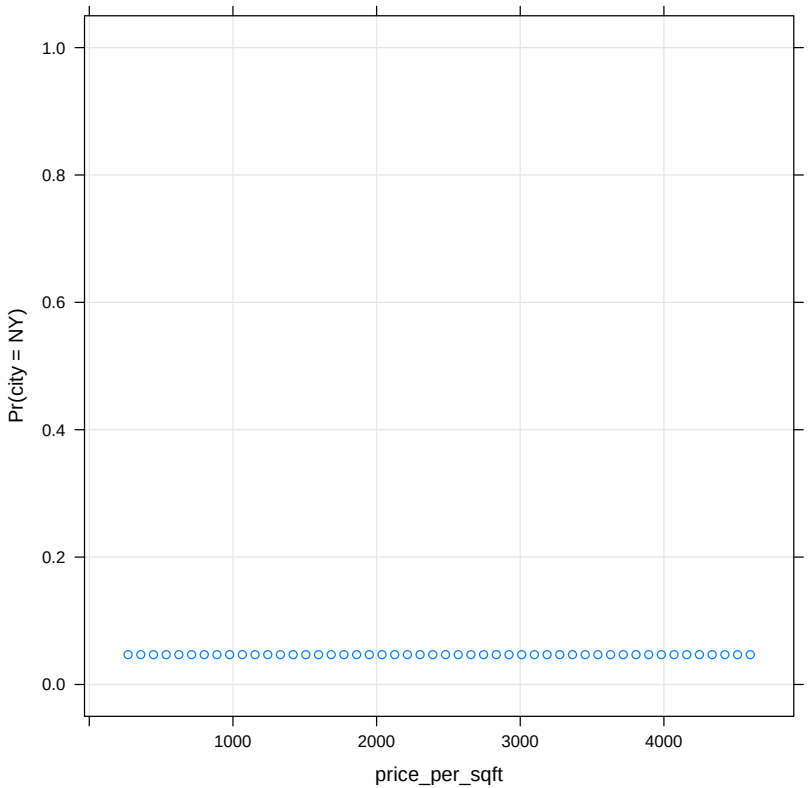
at varying elevation

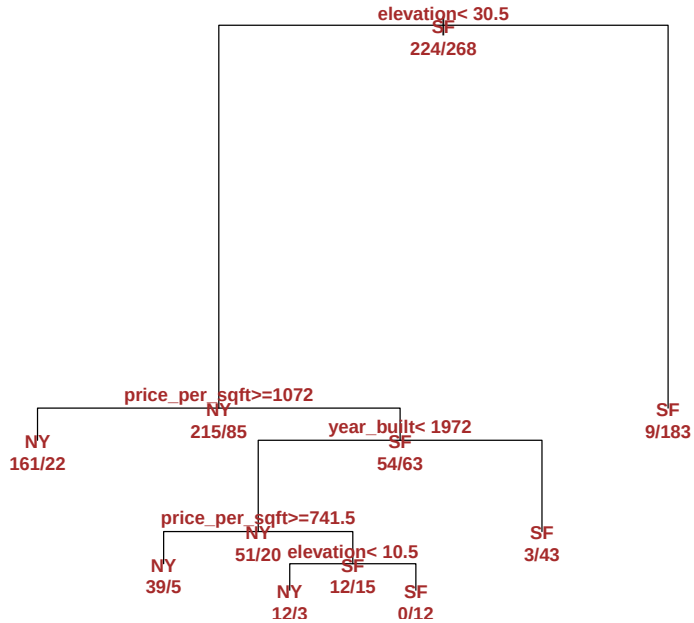




Prediction

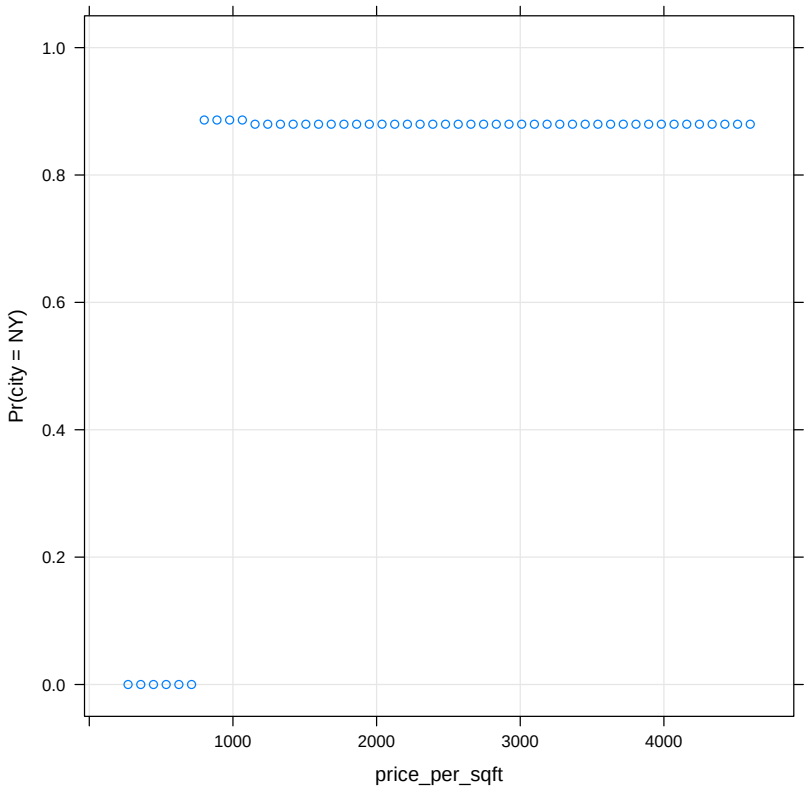
at varying price/sqft and 30m elevation

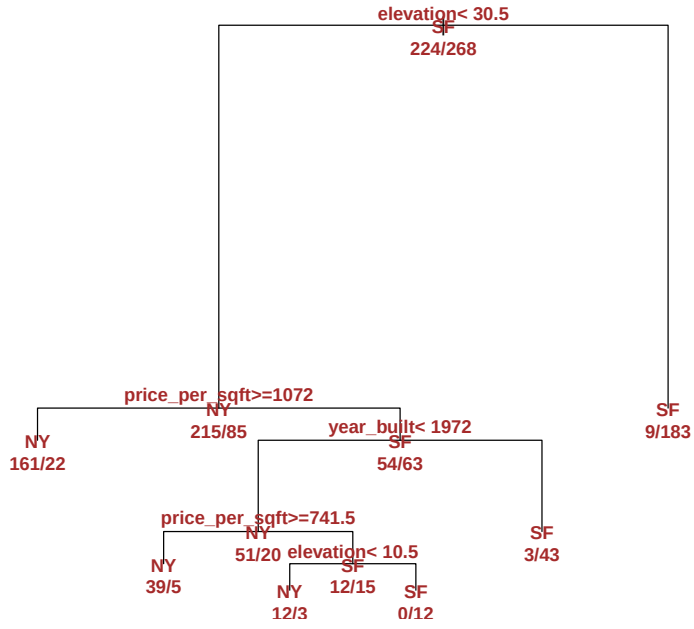




Prediction

at varying price/sqft and average (40m) elevation





Conclusion: Machine Learning and You

1. doing it
   - `Hello World!` is easy but useless
   - obtaining skills takes years
   - rewarding
2. understanding it
   - learn concepts not cooking
   - collaboration, interpretation