Feature Engineering and Selection…
Chapter 1: Introduction

Agenda

1) Navigating materials
2) Preface
3) Introduction
   - Simple Example
   - Important Concepts
   - More Complex Example
   - Feature Selection
   - Outline

Navigating materials

- Online book: http://www.feat.engineering/
- Teams page: LINK
  - See “Files” tab for ppt slides, see “Schedule & recordings links” for links to recordings
  - Feel free to add materials directly! (Do not delete other’s materials)
- Code: https://github.com/topepo/FES (also cloned to Teams page)

Preface

- When you model has poor performance…
- "…lack of performance may be due to a simple to explain, but difficult to pinpoint, cause: relevant predictors that were collected are represented in a way that models have trouble achieving good performance."

- "Adjusting and reworking the predictors to enable models to better uncover predictor-response relationships has been termed **feature engineering**."

Introduction

- 2 uses for a model:
  - Inference/explanation
  - Prediction/estimation
- Parsimony/simplicity of model is usually preferred
  - Easier to trust easier + easier to understand
- HOWEVER, “accuracy should not be seriously sacrificed for simplicity…”
  - Is a balance
  - Help your stakeholders define where this
    balance is in the problem at hand…

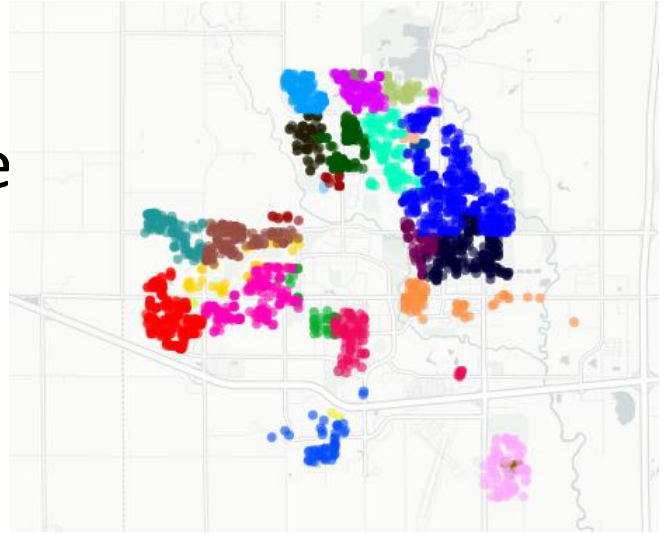
Nomenclature:

Thing you are predicting:

- Target/outcome/response/dependent variable

Thing you use to predict:

- Predictors/features/independent variables/inputs
- “Features” often represent artificial predictors that may be composites or "engineered" versions of initial variables



Target: Sale price of home
Predictors: # bathrooms; # of bedrooms; … ;
 # bathrooms / # bedrooms
How could you encode location?

How might a model for inference be useful?

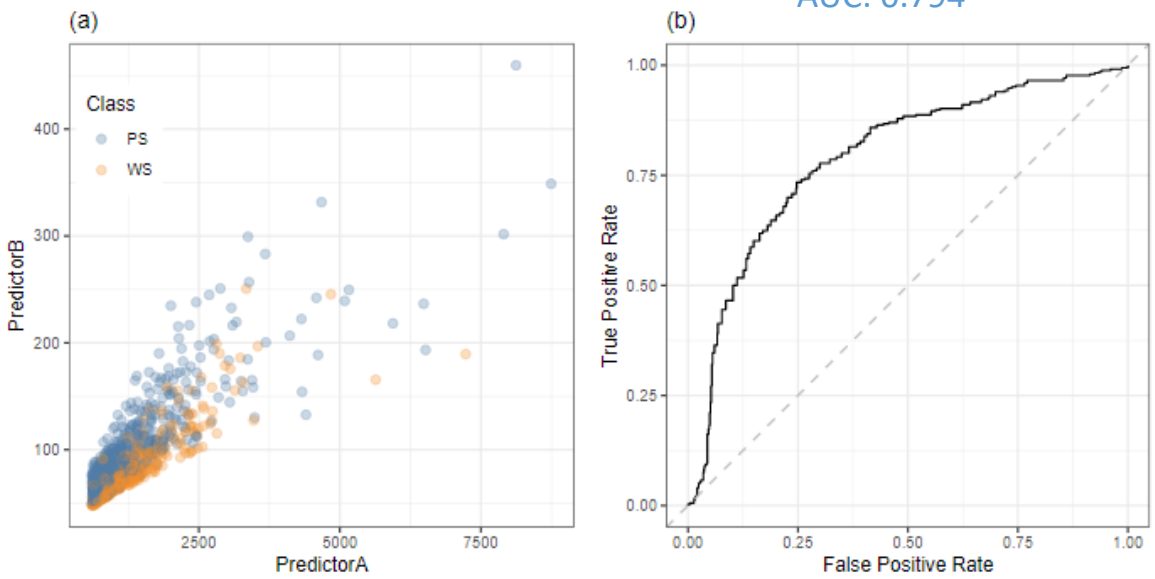
A simple example

What can be some some problems with using prediction accuracy?

- Perform inverse transformation of variables (via Box-cox)

$$log(p/(1-p)) = \beta_0 + \beta_1 A + \beta_2 B$$

AUC: 0.794



$$log(p/(1-p)) = \beta_0 + \beta_1 A + \beta_2 B$$

AUC: 0.848

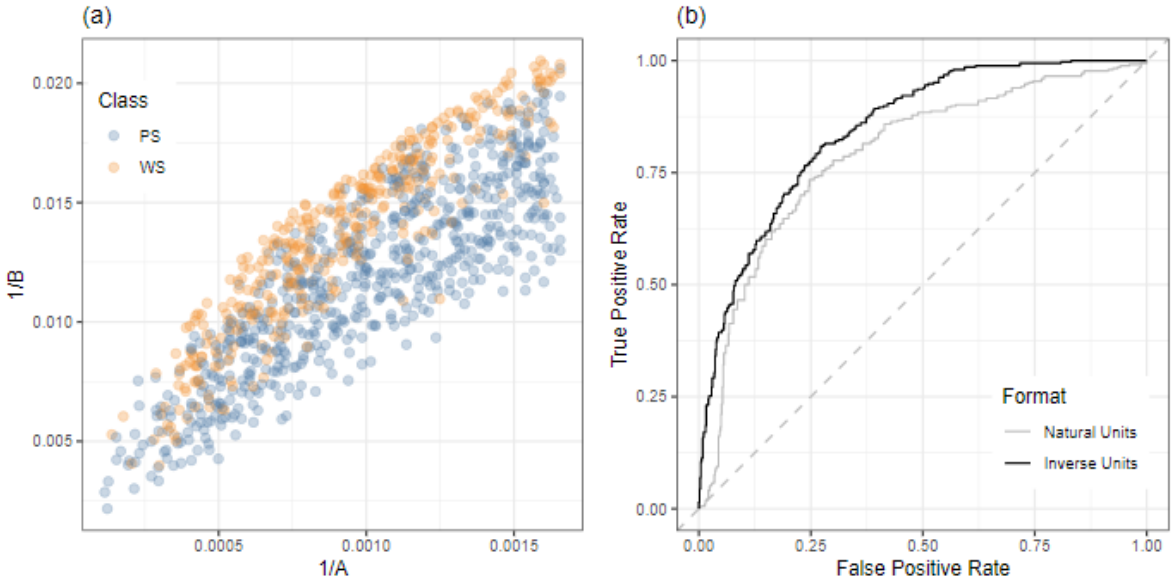


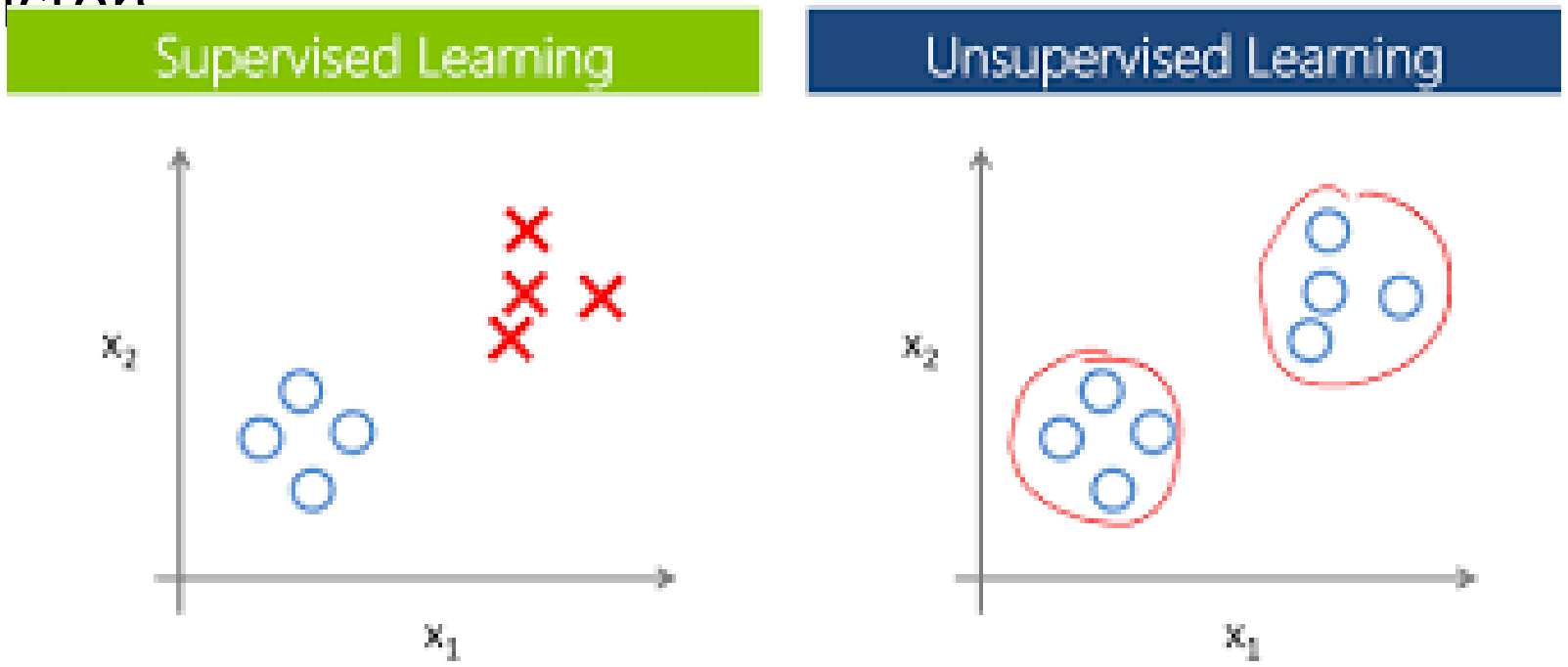
Figure 1.3: (a) The transformed predictors and (b) the ROC curve from both logistic regression models.

Neural network to this model, which is not necessarily sensitive to this skewness, will get an AUC roughly equivalent to logistic regression (after transformation)… neural nets have other drawbacks

Important concepts

Supervised and Unsupervised Procedures

- “Supervised data analysis involves identifying patterns between predictors and an identified *outcome* that is to be modeled or predicted, while unsupervised techniques are focused solely on identifying patterns among the predictors.”



https://lakshaysuri.wordpress.com/2017/03/19/machine-learning-supervised-vs-unsupervised-learning/

Important concepts

No Free Lunch

- “Without any specific knowledge of the problem or data at hand, no one predictive model can be said to be the best.”

- “It is wise to try a number of disparate types of models to probe which ones will work well with your particular data set.”

Important concepts

Should you split your data into a training dataset before or after performing exploratory data analysis?

The model versus the modeling process

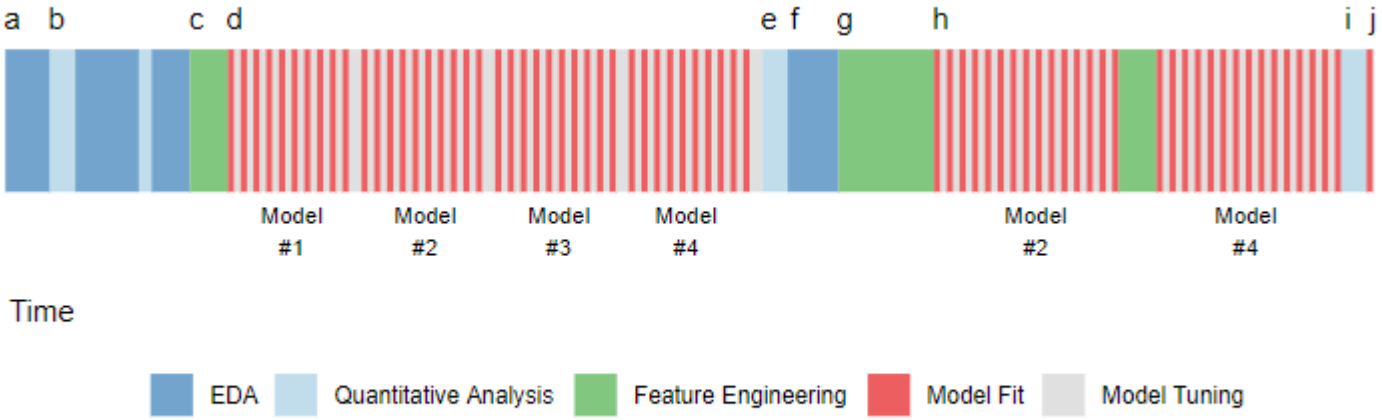
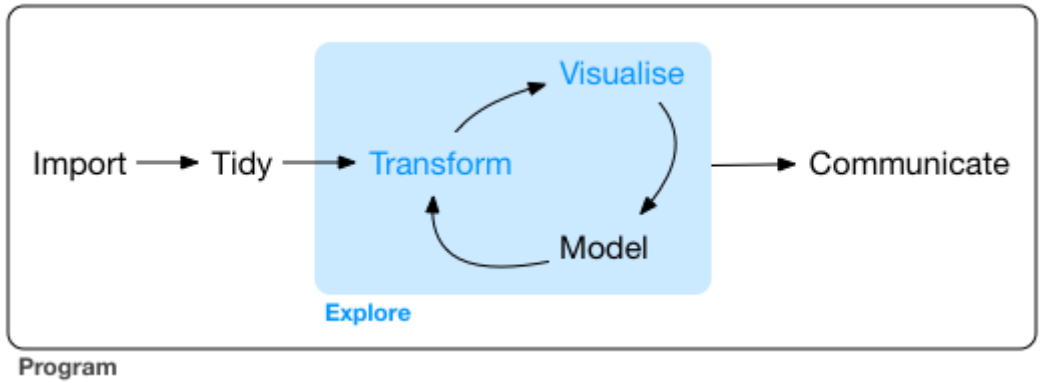


Figure 1.4: A schematic for the typical modeling process.



https://r4ds.had.co.nz/

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)

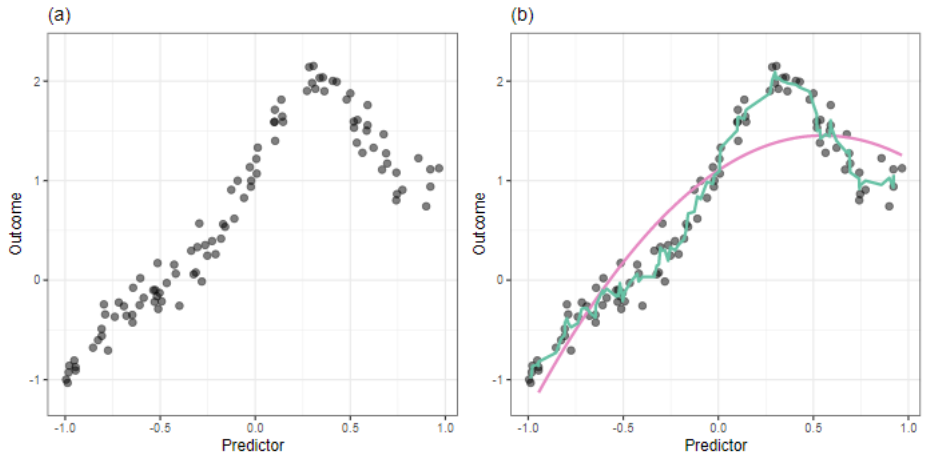


Figure 1.5: A simulated data set and model fits for a 3-point moving average (green) and quadratic regression (purple).

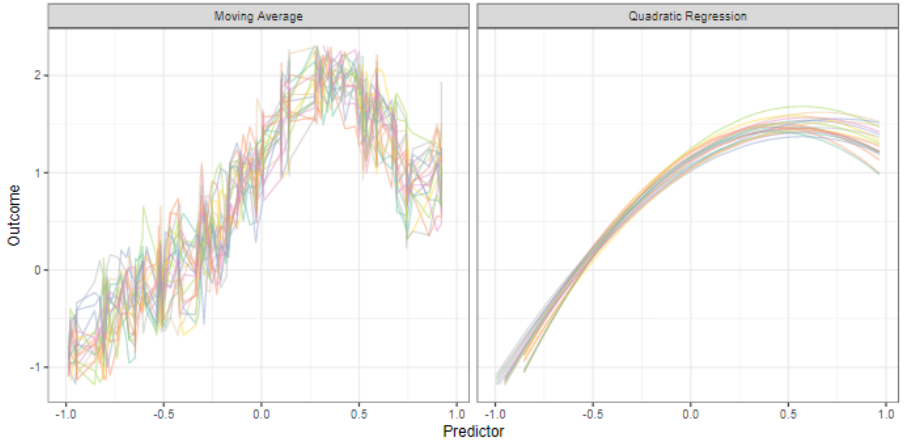
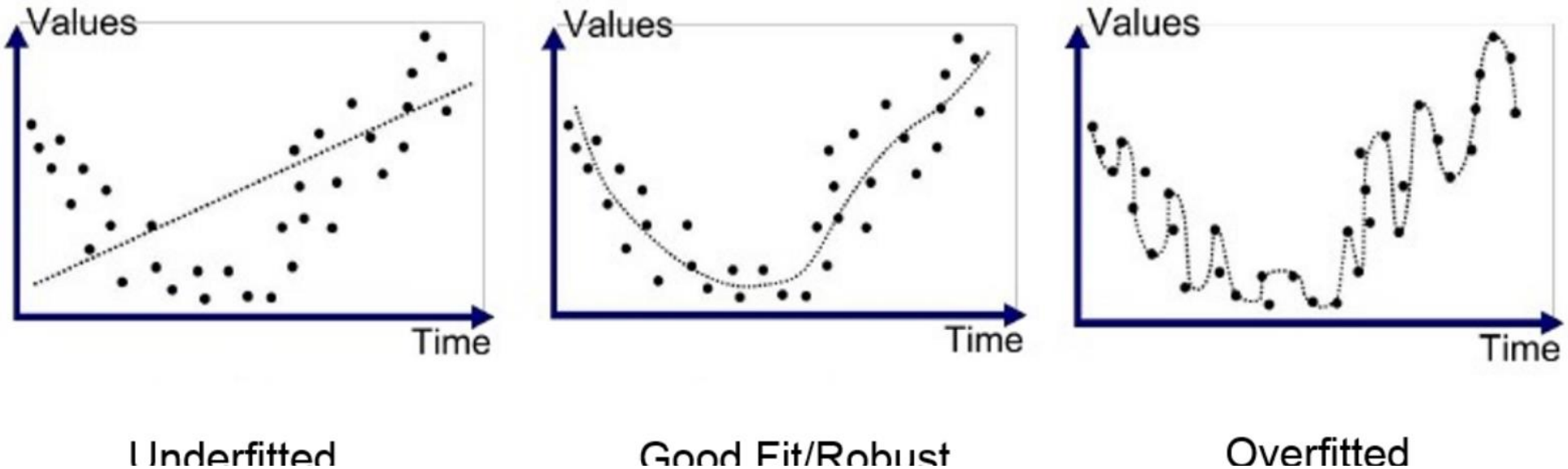


Figure 1.6: Model fits for twenty jittered versions of the data set.

Important concepts

Underfitting and Overfitting



Underfitted Good Fit/Robust Overfitted

https://medium.com/greyatom/what-is-underfitting-and-overfitting-in-machine-learning-and-how-to-deal-with-it-6803a989c76

A more complex example

What are other factors to consider other than performance?

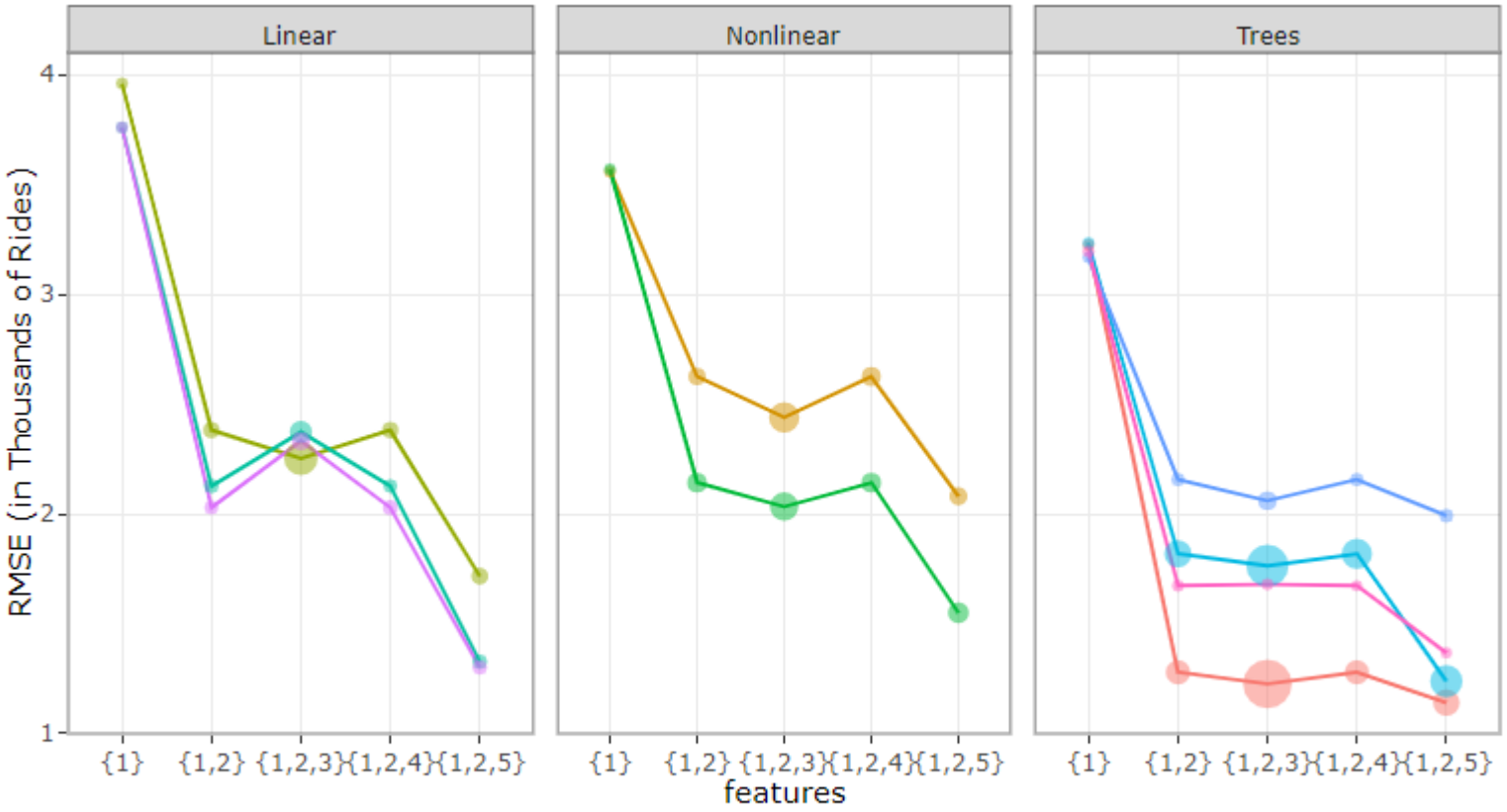


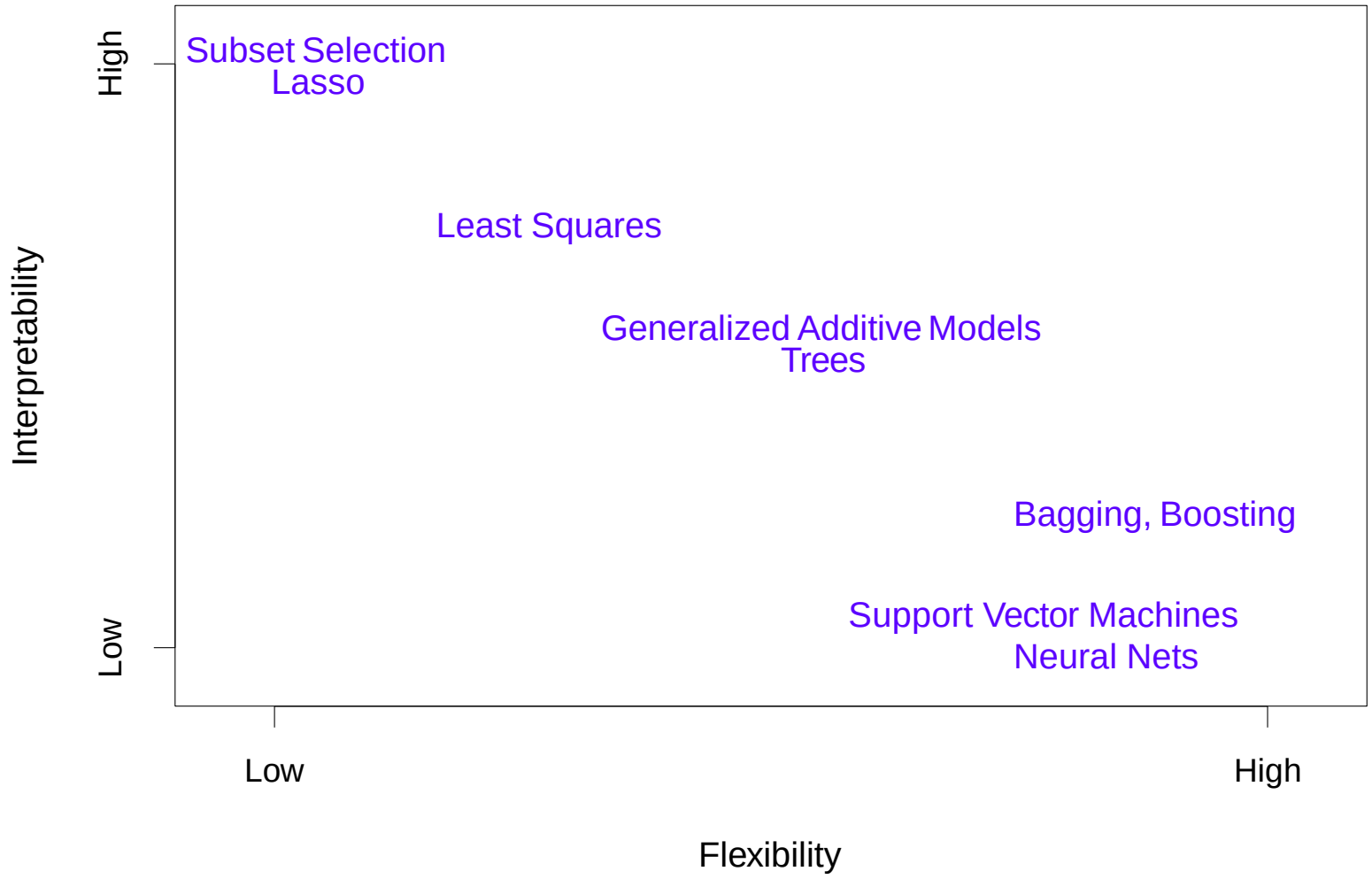
Figure 1.7: A series of model and feature set combinations for modeling train ridership in Chicago.

1. When modeling data, there is almost never a single model fit or feature set that will immediately solve the problem. The process is more likely to be a *campaign* of trial and error to achieve the best results.
2. The effect of feature sets *can* be much larger than the effect of different models.
3. The interplay between models and features is complex and somewhat unpredictable.
4. With the right set of predictors, is it common that many different types of models can achieve the same level of performance. Initially, the linear models had the worst performance but, in the end, showed some of the best performance

Feature Selection

- Wrapper methods
  - E.g. stepwise / recursive feature selection techniques; genetic algorithms
- Embedded methods
  - E.g. splits in decision trees; lasso penalty
- Unsupervised selection
  - E.g. remove rare levels if dummy encoding

Model types by interpretability vs flexibility



Introduction to Statistical Learning with Applications in R, http://faculty.marshall.usc.edu/gareth-james/ISL/