Feature Engineering

Chapter 3 Part 2

Stephen Kimel

Splitting Data

- "The training set is used to develop models and feature sets…The **test** set is used only at the conclusion of these activities for estimating a final, unbiased assessment of the model's performance. It is **critical** that the **test** set not be used prior to this point."
- How much should be set aside for testing?

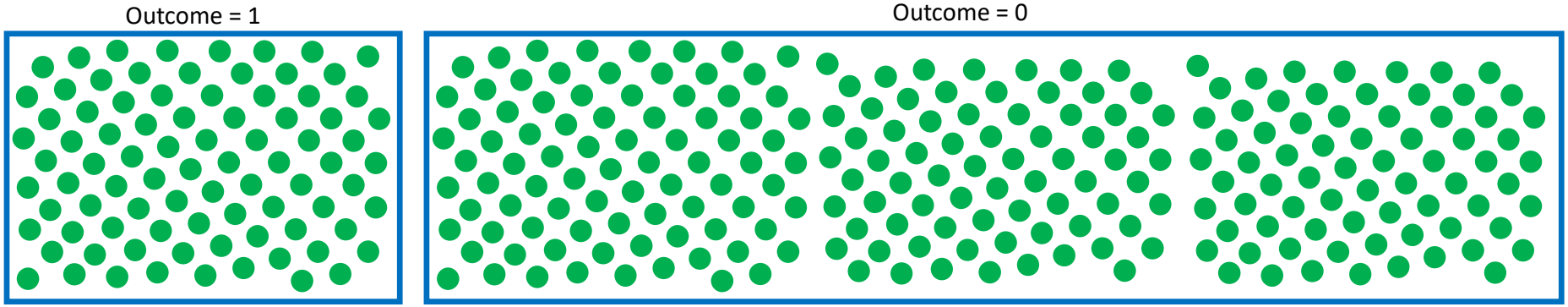
Observation #	X-variable p1	X-variable p2	X-variable p3	X-variable p4	X-variable p5
n1					
…					
n100,000					

Observation #	X-variable p1	X-variable p2	X-variable p3	…	X-variable p1000
n1					
…					
n100					

Sampling Methods

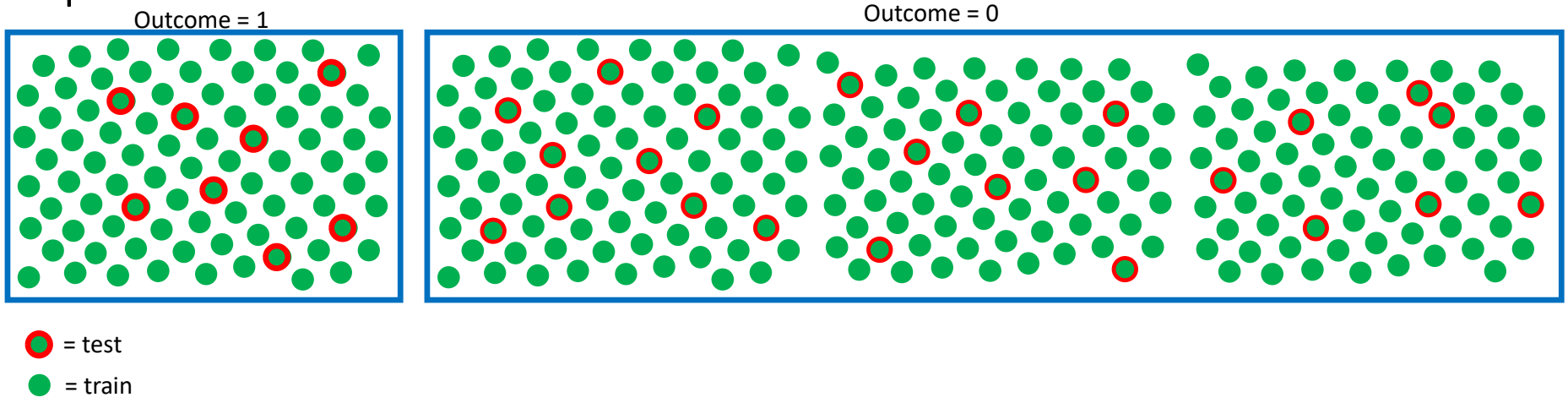
Determining which observations go into which buckets

- Random
- Stratified
  - “For classification models, this is accomplished by selecting samples at random *within* each class.”
  - “When the outcome is numeric, artificial strata can be constructed based on the quartiles of the data.”

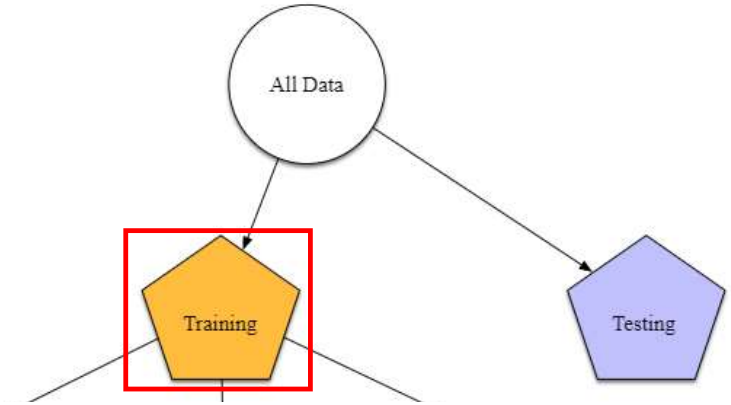


Sampling Methods

- Random
- Stratified
  - “For classification models, this is accomplished by selecting samples at random *within* each class.”
  - “When the outcome is numeric, artificial strata can be constructed based on the quartiles of the data.”



Resampling



V-Fold (or K-Fold) Cross-Validation

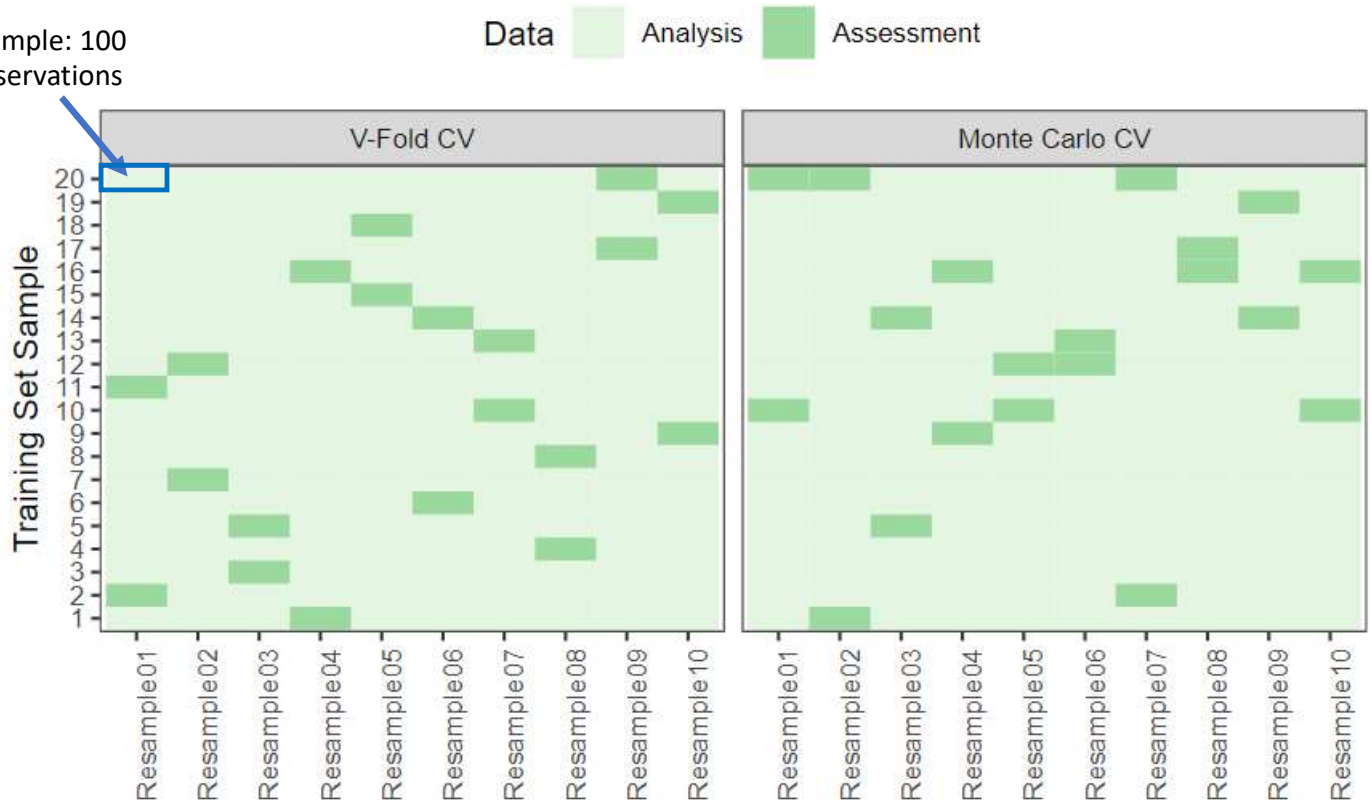
Divide data into K roughly equal-sized parts ($K = 5$ here)

1	2	3	4	5
Assessment1 RMSE	Analysis	Analysis	Analysis	Analysis
Analysis	Assessment2 RMSE	Analysis	Analysis	Analysis
Analysis	Analysis	Assessment3 RMSE	Analysis	Analysis
Analysis	Analysis	Analysis	Assessment4 RMSE	Analysis
Analysis	Analysis	Analysis	Analysis	Assessment5 RMSE

Overall Model Performance = mean(Assessment1 RMSE…Assessment5 RMSE)

Cross-Validation

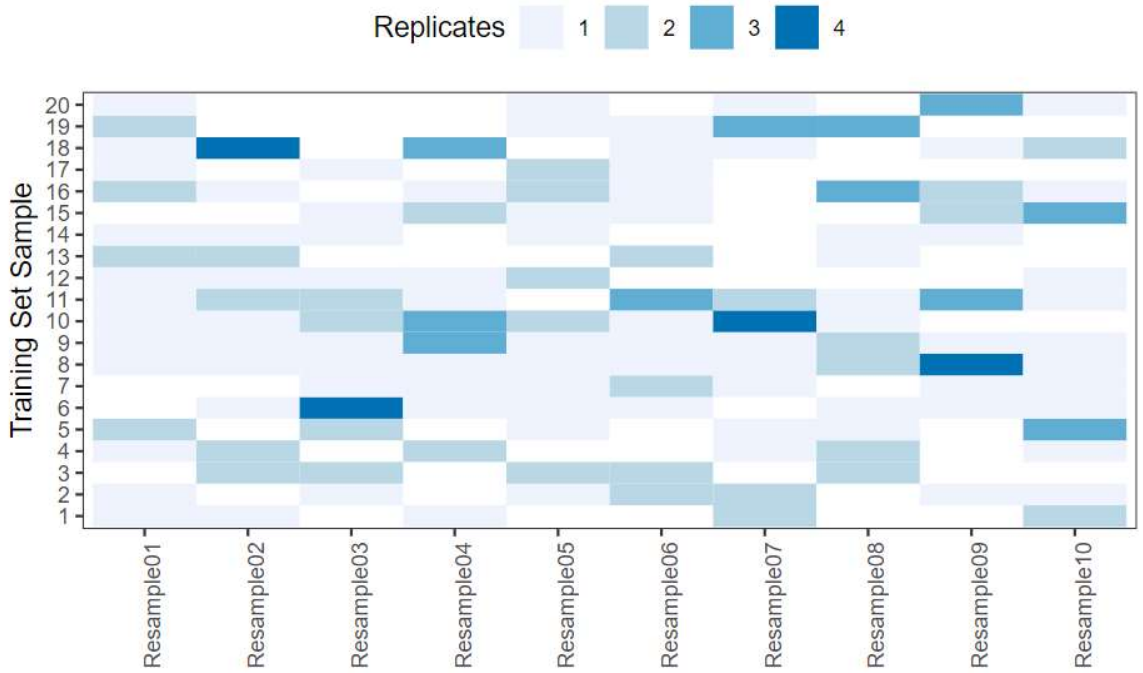
Example: 100 Observations



NOTE: For customer transactional data, it is recommended to keep all transactions of a customers entirely in the Analysis or Assessment group.

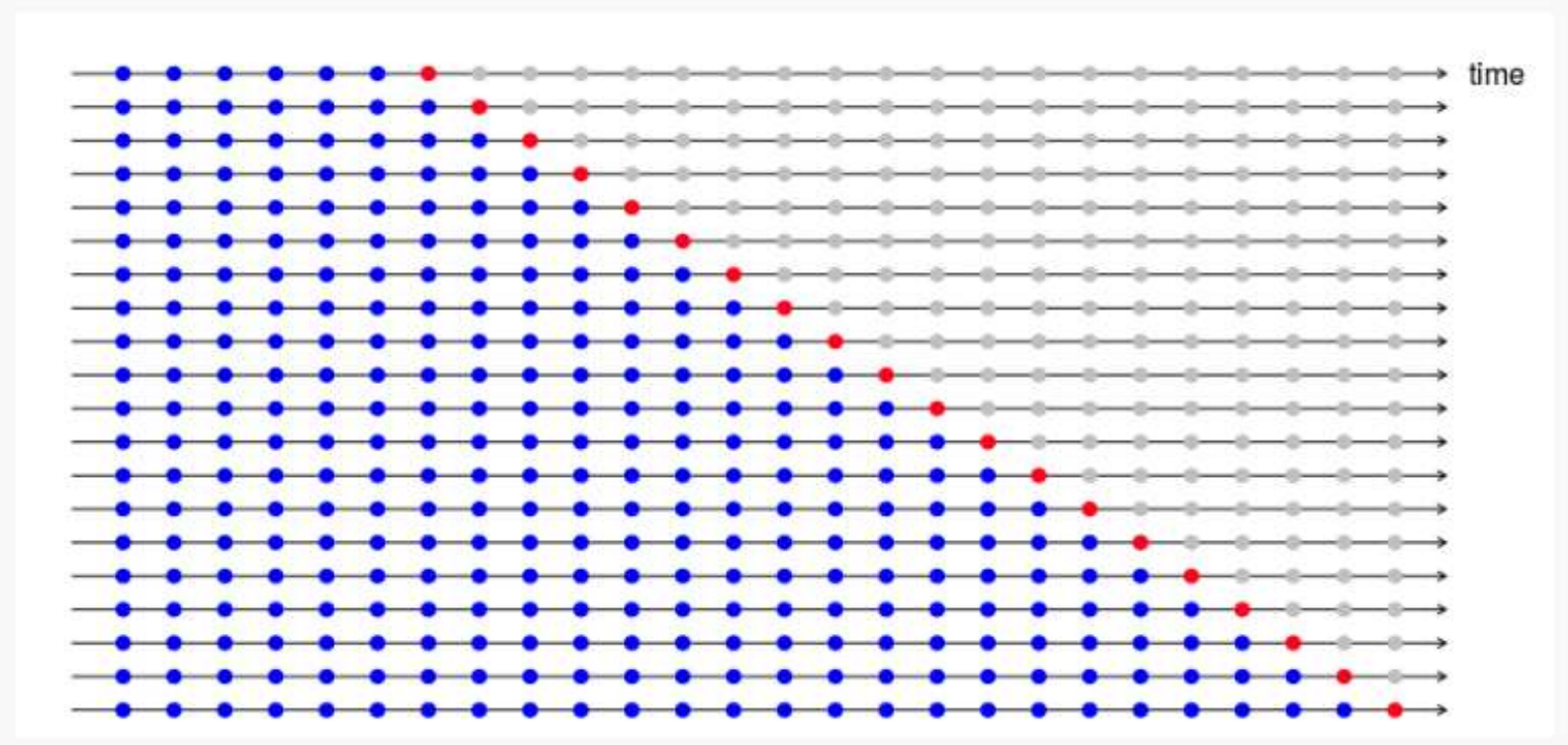
The Bootstrap

A bootstrap resample of the data is defined to be a simple random sample that is the same size as the training set where the data are sampled *with replacement*



Rolling Forecasting

Forecasting Assessment (Time Series)



Source: https://robjhyndman.com/hyndsight/tscv/

Important concepts

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)

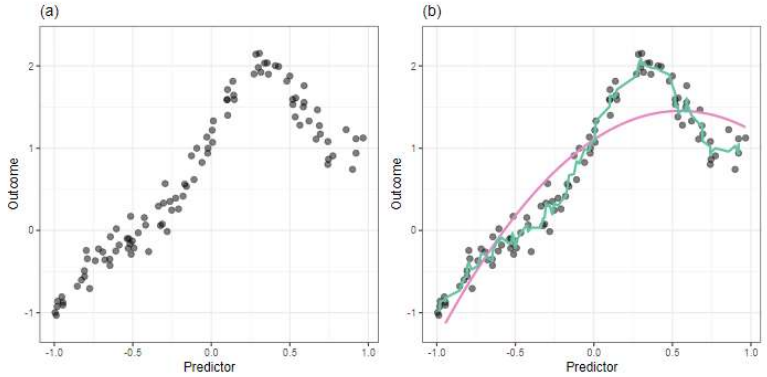


Figure 1.5: A simulated data set and model fits for a 3-point moving average (green) and quadratic regression (purple).

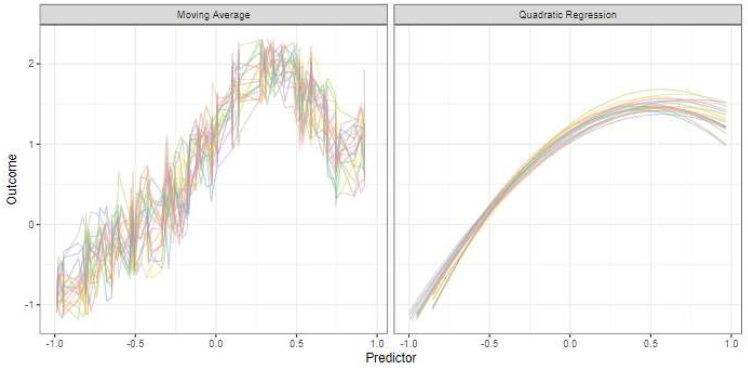


Figure 1.6: Model fits for twenty jittered versions of the data set.

Bias

- "**Bias** is the ability of a particular resampling scheme to be able to hit the **true** underlying performance *parameter* (that we will never truly know). Generally speaking, as the amount of **data** in the analysis set **shrinks**, the resampling estimate's **bias increases**. In other words, the bias in 10-fold cross-validation is smaller than the bias in 5-fold cross-validation."

Variance

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

1	2	3	4	5
Assessment1 RMSE	Analysis	Analysis	Analysis	Analysis
Analysis	Assessment2 RMSE	Analysis	Analysis	Analysis
Analysis	Analysis	Assessment3 RMSE	Analysis	Analysis
Analysis	Analysis	Analysis	Assessment4 RMSE	Analysis
Analysis	Analysis	Analysis	Analysis	Assessment5 RMSE

Overall Model Performance = mean(Assessment1 RMSE…Assessment5 RMSE)

Variance

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

K1	K2	K3	K 4	K5
Ob1 Ob2 Ob3 Ob4	Ob1 Ob2 Ob3 Ob4	Ob1 Ob2 Ob3 Ob4	Ob1 Ob2 Ob3 Ob4	Ob1 Ob2 Ob3 Ob4
Ob5 Ob6 Ob7 Ob8	Ob5 Ob6 Ob7 Ob8	Ob5 Ob6 Ob7 Ob8	Ob5 Ob6 Ob7 Ob8	Ob5 Ob6 Ob7 Ob8
Ob9 Ob10 Ob11 Ob12	Ob9 Ob10 Ob11 Ob12	Ob9 Ob10 Ob11 Ob12	Ob9 Ob10 Ob11 Ob12	Ob9 Ob10 Ob11 Ob12
Ob13 Ob14 Ob15 Ob16	Ob13 Ob14 Ob15 Ob16	Ob13 Ob14 Ob15 Ob16	Ob13 Ob14 Ob15 Ob16	Ob13 Ob14 Ob15 Ob16
Ob17 Ob18 Ob19 Ob20	Ob17 Ob18 Ob19 Ob20	Ob17 Ob18 Ob19 Ob20	Ob17 Ob18 Ob19 Ob20	Ob17 Ob18 Ob19 Ob20

Overall Model Performance = mean(Assessment1 RMSE…Assessment5 RMSE)

OMP1 = 5.5

Variance

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

K1	K2	K3	K 4	K5
Ob1 Ob12 Ob10 Ob20	Ob1 Ob12 Ob10 Ob20	Ob1 Ob12 Ob10 Ob20	Ob1 Ob12 Ob10 Ob20	Ob1 Ob12 Ob10 Ob20
Ob5 Ob9 Ob11 Ob18	Ob5 Ob9 Ob11 Ob18	Ob5 Ob9 Ob11 Ob18	Ob5 Ob9 Ob11 Ob18	Ob5 Ob9 Ob11 Ob18
Ob16 Ob2 Ob3 Ob6	Ob16 Ob2 Ob3 Ob6	Ob16 Ob2 Ob3 Ob6	Ob16 Ob2 Ob3 Ob6	Ob16 Ob2 Ob3 Ob6
Ob13 Ob17 Ob4 Ob19	Ob13 Ob17 Ob4 Ob19	Ob13 Ob17 Ob4 Ob19	Ob13 Ob17 Ob4 Ob19	Ob13 Ob17 Ob4 Ob19
Ob7 Ob8 Ob14 Ob15	Ob7 Ob8 Ob14 Ob15	Ob7 Ob8 Ob14 Ob15	Ob7 Ob8 Ob14 Ob15	Ob7 Ob8 Ob14 Ob15

Overall Model Performance = mean(Assessment1 RMSE…Assessment5 RMSE)

OMP1 = 5.5 OMP2 = 5.9 OMP3 = 4.2 … OMP100 = 5.0

Preprocessing

- Preprocess the data for each iteration
  - Example: When imputing values for K1 with the median, use the median from the data in the Analysis set of data

Information Leakage

- Can you predict the while observations are in the Analysis and Assessment buckets?
- “Another, more overt path to information leakage, can sometimes be seen in machine learning competitions (Kaggle) where the training and test set data are given at the same time. While the test set data often have the outcome data blinded, it is possible to “train to the test” by only using the training set samples that are most similar to the test set data.”

Tuning Parameters (Hyperparameters) and Overfitting

- Hyperparameters are model-specific knobs to turn
  - i.e. number of trees in a random forest or how many variables to consider at each split
- Bias-Variance trade-off

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