Data Usage

Matthew McDonald

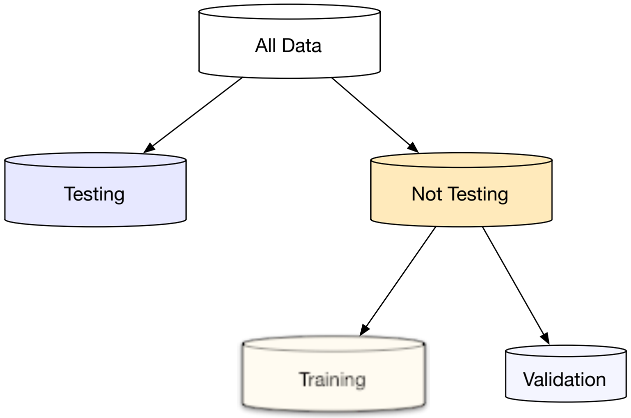
Data splitting and spending

For machine learning, we typically split data into training and test sets:

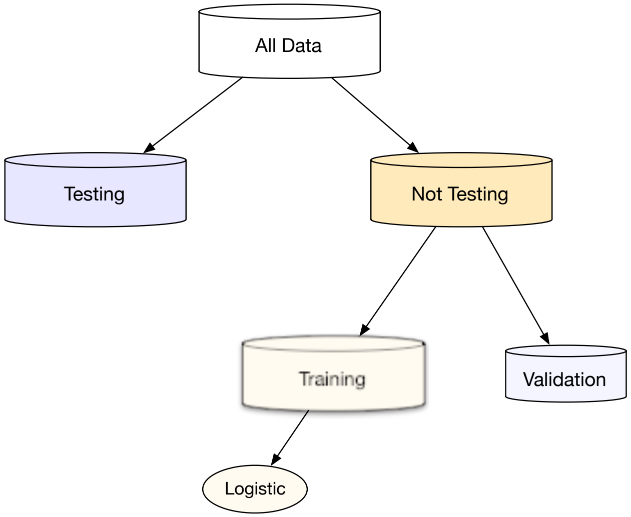
- The **training set** is used to estimate model parameters.
- The **test set** is used to find an independent assessment of model performance.

Do not 🚫 use the test set during training.

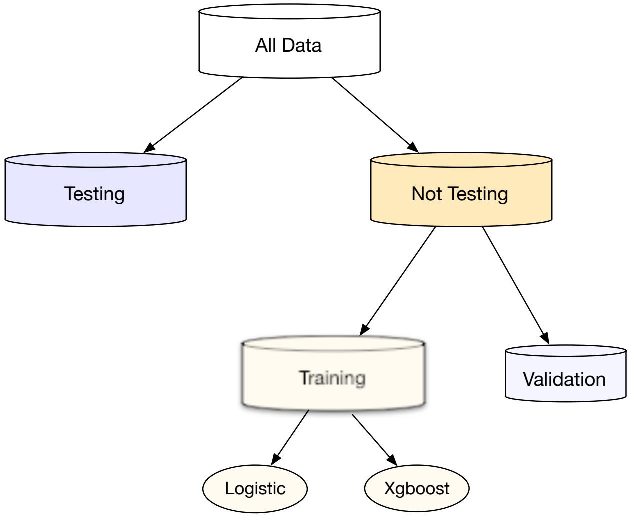
Data spending



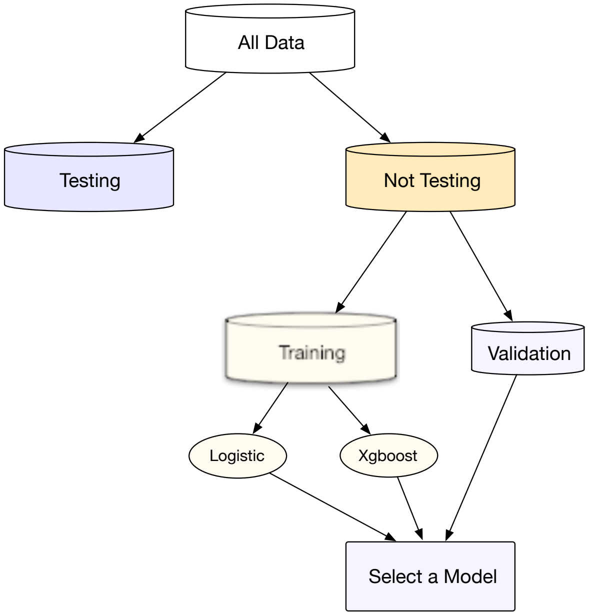
A first model



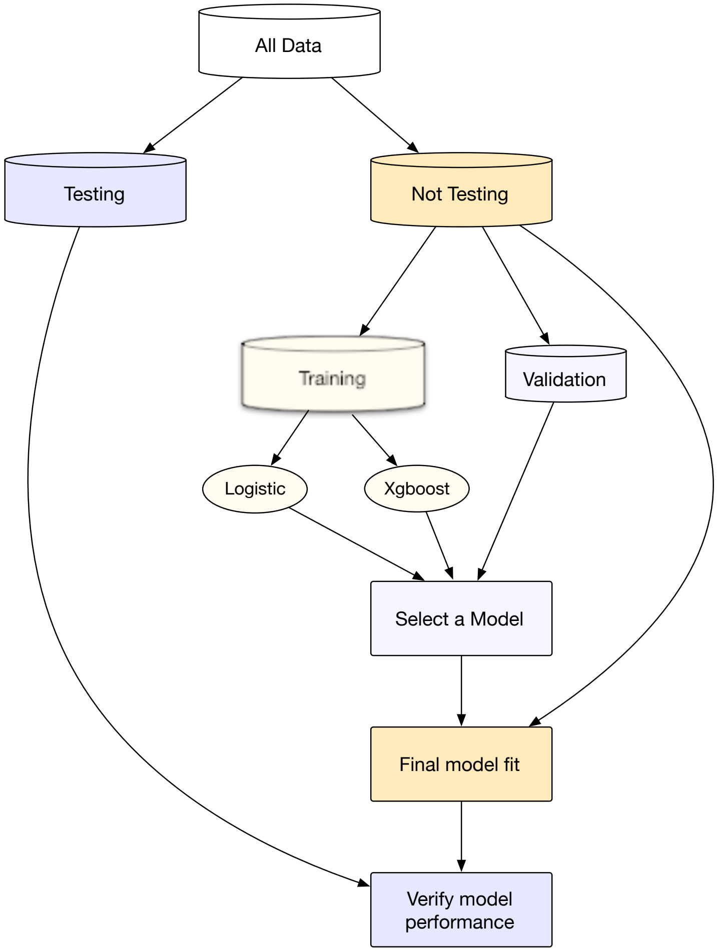
Try another model



Choose wisely…



Finalize and verify



… and so on

Once we find an acceptable model and feature set, the process is to

- Confirm our results on the test set.
- Document the data and model development process.
- Deploy, monitor, etc.

Data splitting and spending

- Spending too much data in **training** prevents us from computing a good assessment of predictive **performance**.

- Spending too much data in **testing** prevents us from computing a good estimate of model **parameters**.

How Do We Split THe Data?

```r
library(tidymodels)
tidymodels_prefer()

# Set the random number stream using `set.seed()` so that the results
# can be reproduced later.
set.seed(501)

# Save the split information for an 80/20 split of the data
ames_split <- initial_split(ames, prop = 0.80)
ames_split
```

```
<Training/Testing/Total>
<2344/586/2930>
```

Getting the Resulting Dataframes

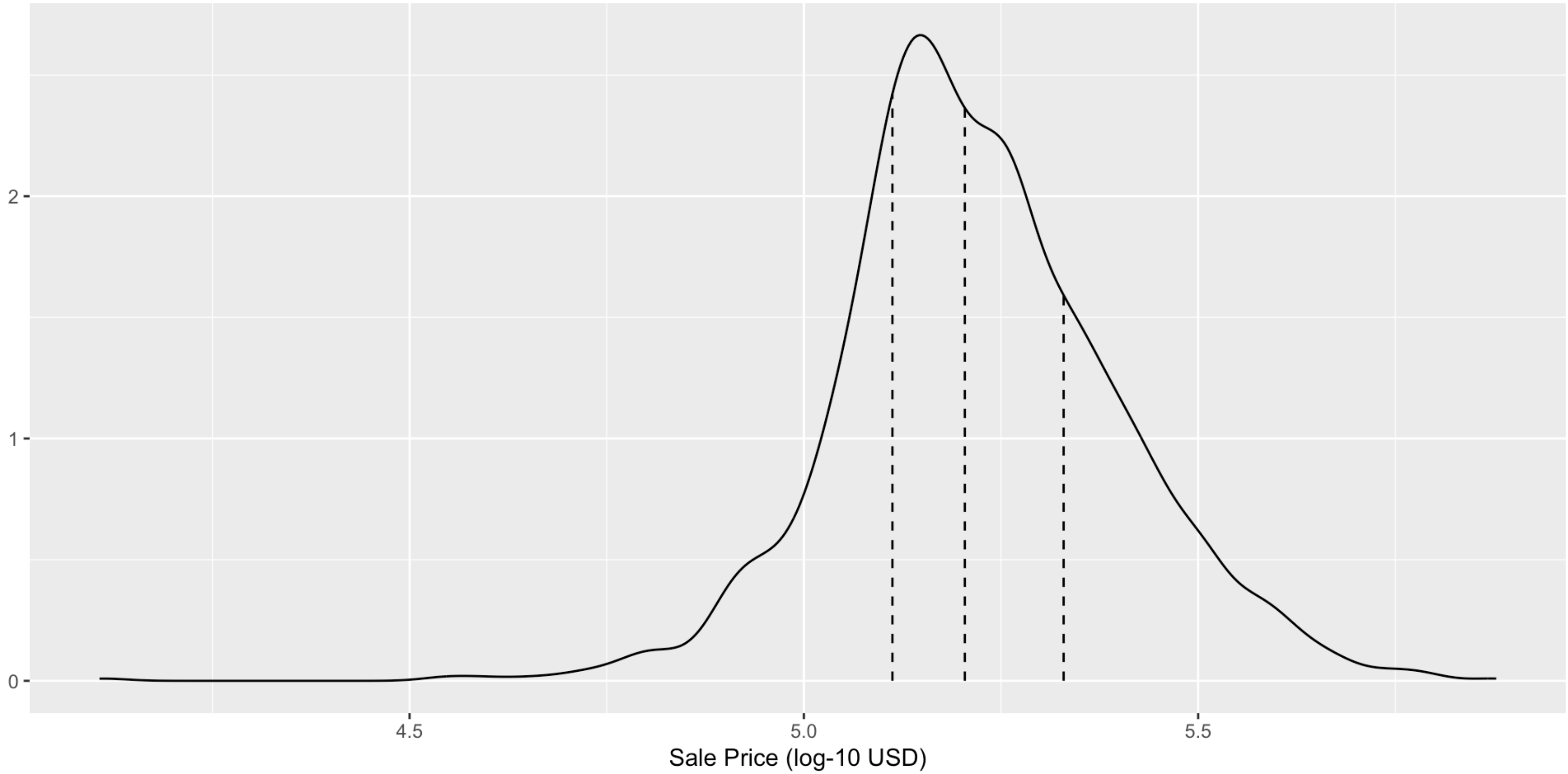
```r
ames_train <- training(ames_split)
ames_test  <-  testing(ames_split)

dim(ames_train)
```

```
[1] 2344   74
```

Stratified Sampling

- Simple random sampling is appropriate in many cases but there are exceptions.
- When there is a dramatic *class imbalance* in classification problems, one class occurs much less frequently than another.
- Using a simple random sample may haphazardly allocate these infrequent samples disproportionately into the training or test set.
- To avoid this, *stratified sampling* can be used.
- The training/test split is conducted separately within each class and then these subsamples are combined into the overall training and test set.
- For regression problems, the outcome data can be artificially binned into quartiles and then stratified sampling can be conducted four separate times.

Ames Sale Price



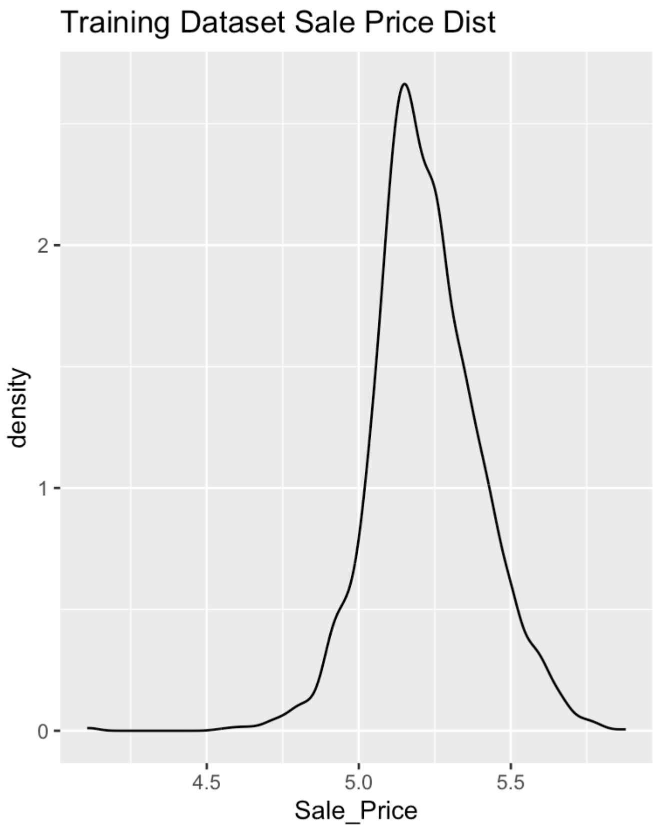
The distribution of the sale price (in log units) for the Ames housing data. The vertical lines indicate the quartiles of the data

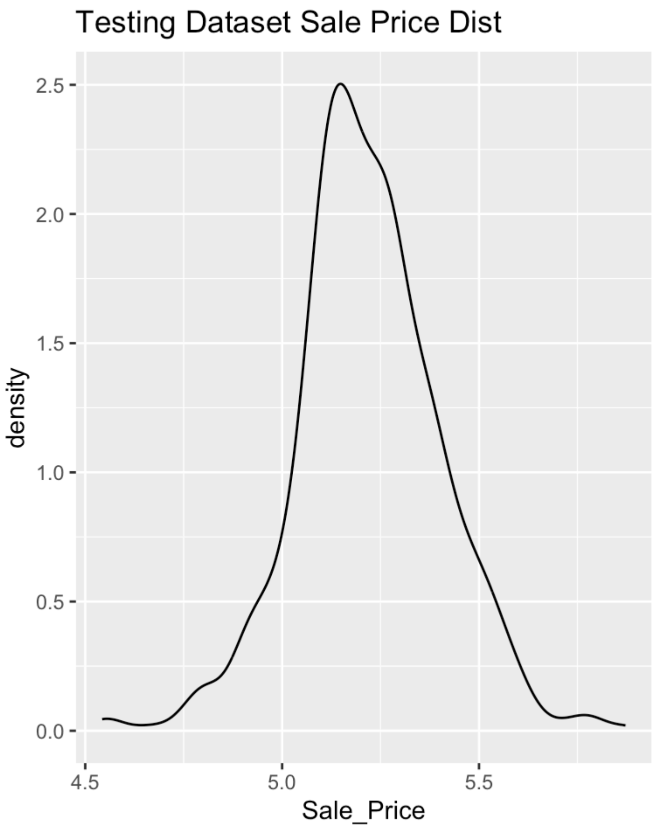
Creating A Stratified Split

```r
set.seed(502)
ames_split <- initial_split(ames, prop = 0.80, strata = Sale_Price)
ames_train <- training(ames_split)
ames_test  <-  testing(ames_split)

dim(ames_train)
```

```
[1] 2342   74
```

Resulting Distributions of Sale Price


Training Dataset Sale Price Dist


Testing Dataset Sale Price Dist