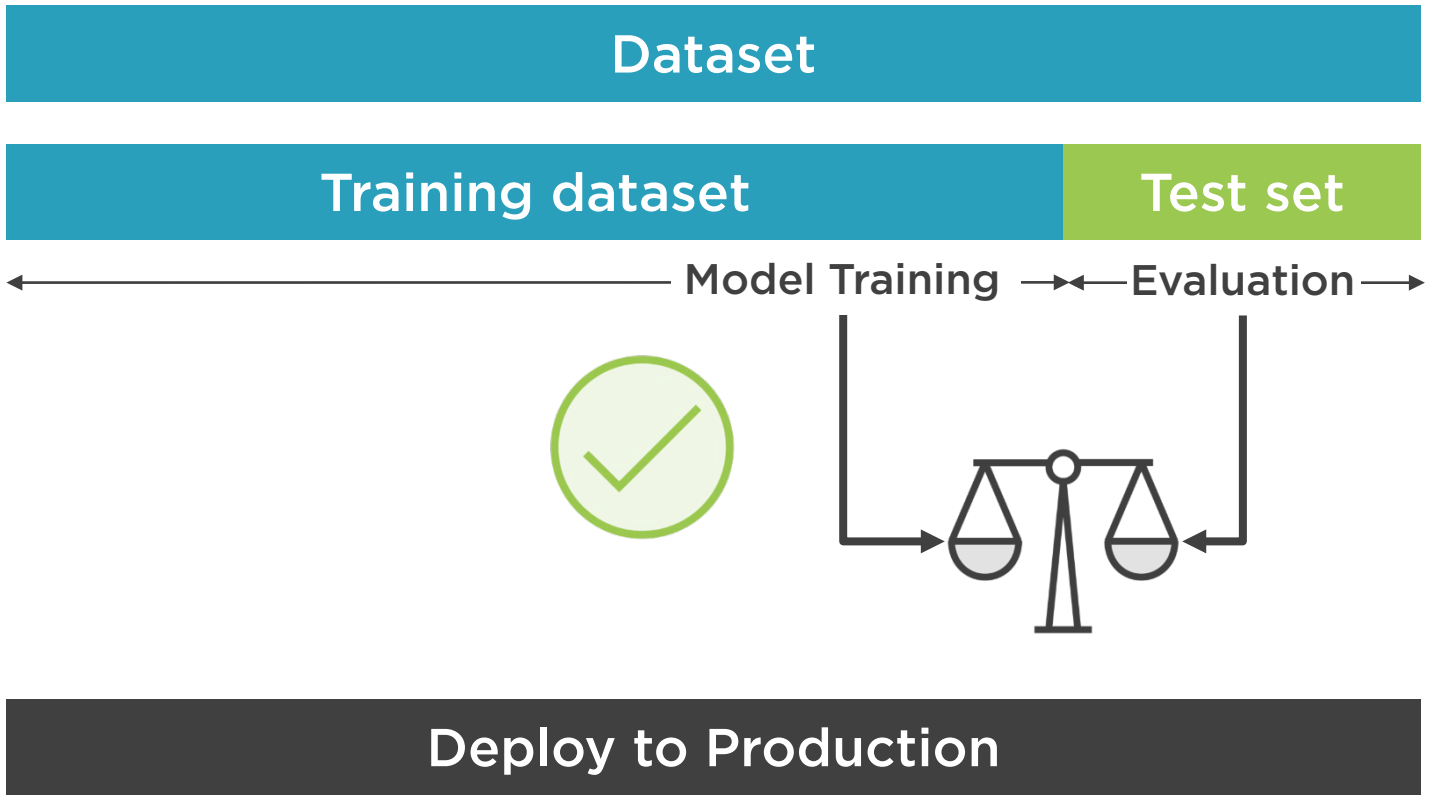
Split a Dataset into Training and Test Subsets

Ravikiran Srinivasulu
SOFTWARE CONSULTANT

ravikirans.com | ravikirans.com/YouTube

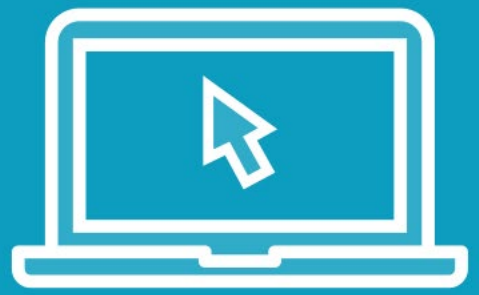
Why Split the Data in Machine Learning?



Agenda

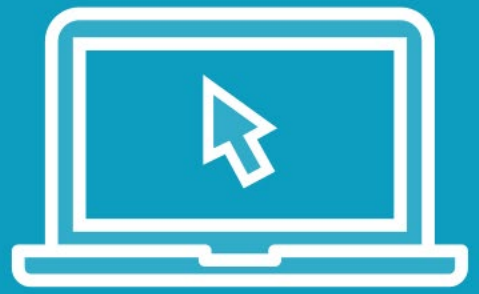
- Model training and evaluation on same data
- Split the data into training and test set
- Split the data for Model tuning
- Cross-validation
- Model selection

Demo



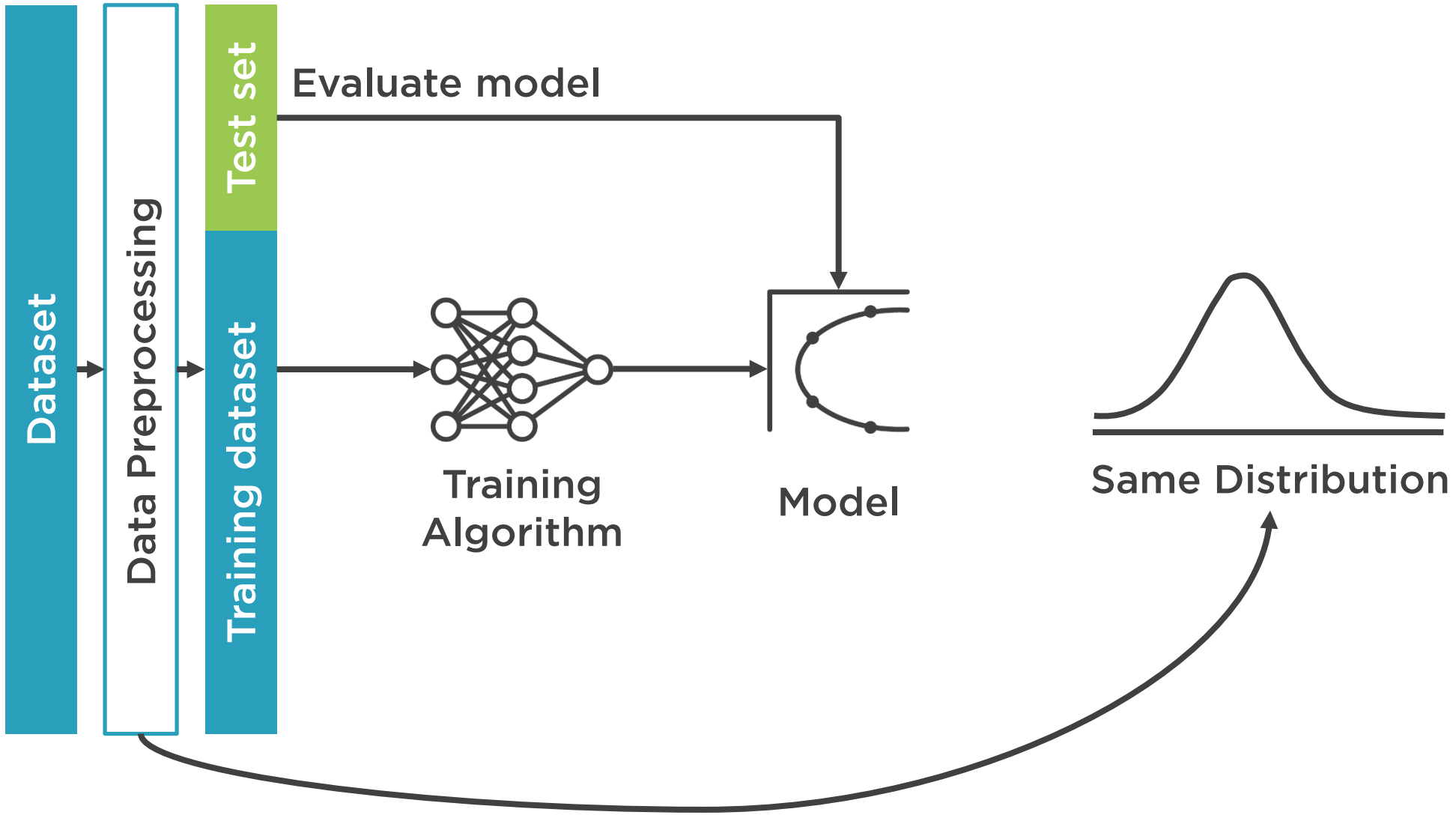
Training and testing on same data

Demo

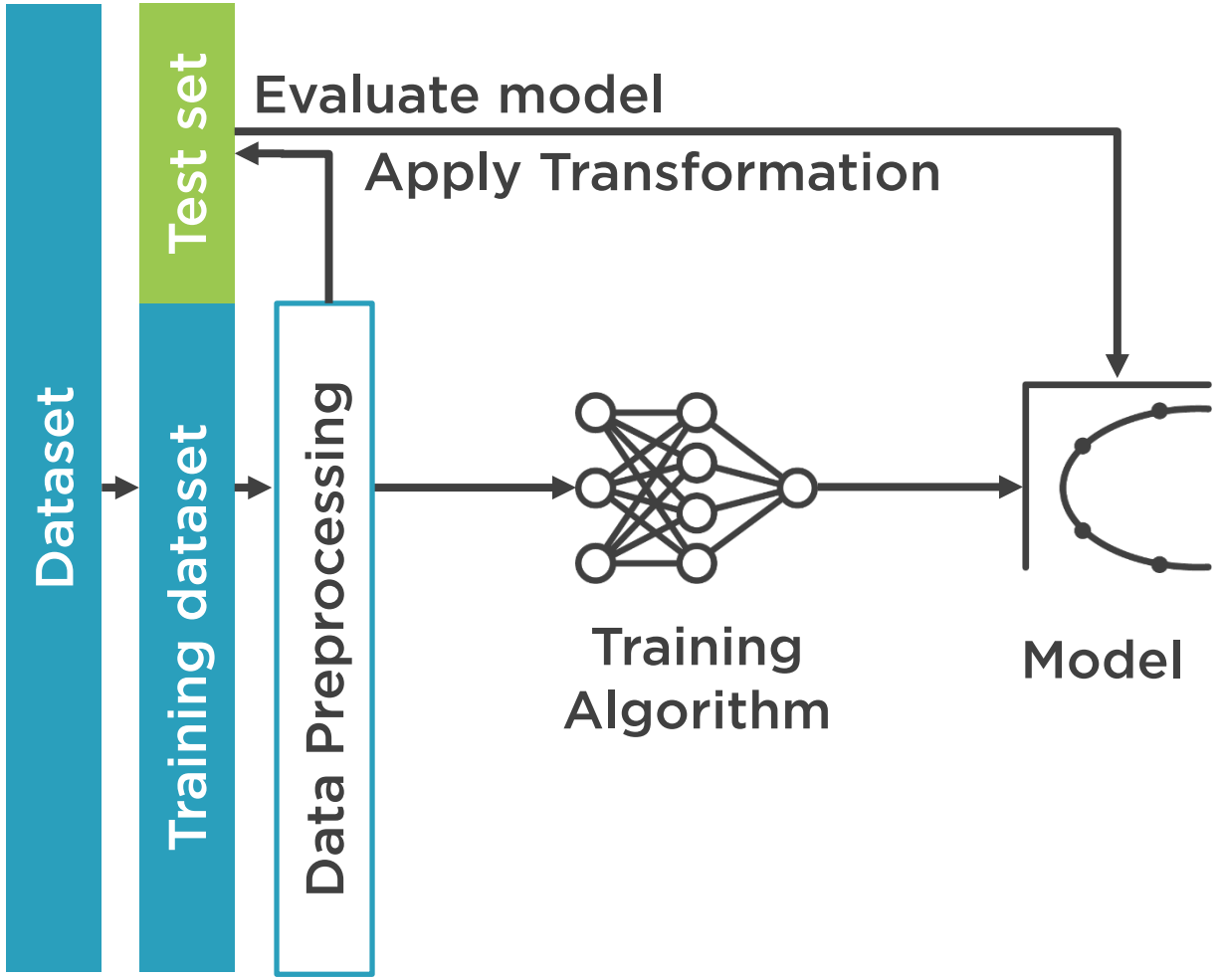


Split data into training and testing set

Data Leakage



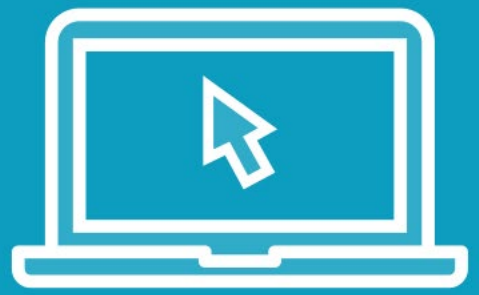
Better Way


Dataset
Test set
Training dataset
Data Preprocessing
Evaluate model
Apply Transformation
Training Algorithm
Model

Disadvantages of Train-test-split

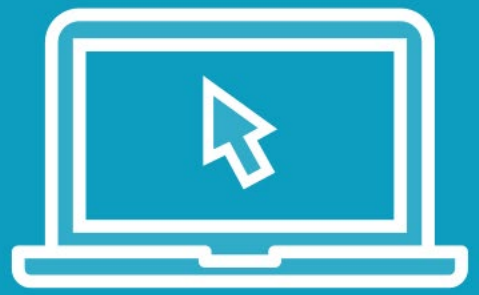
- The model loses available data to learn from
- Test metrics vary a lot depending on how data is split

Demo



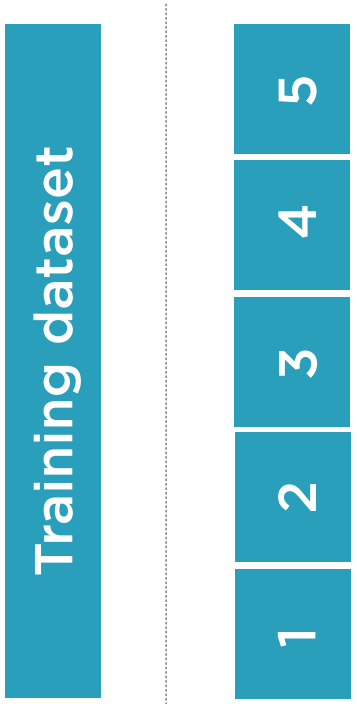
Split dataset for tuning models

Demo

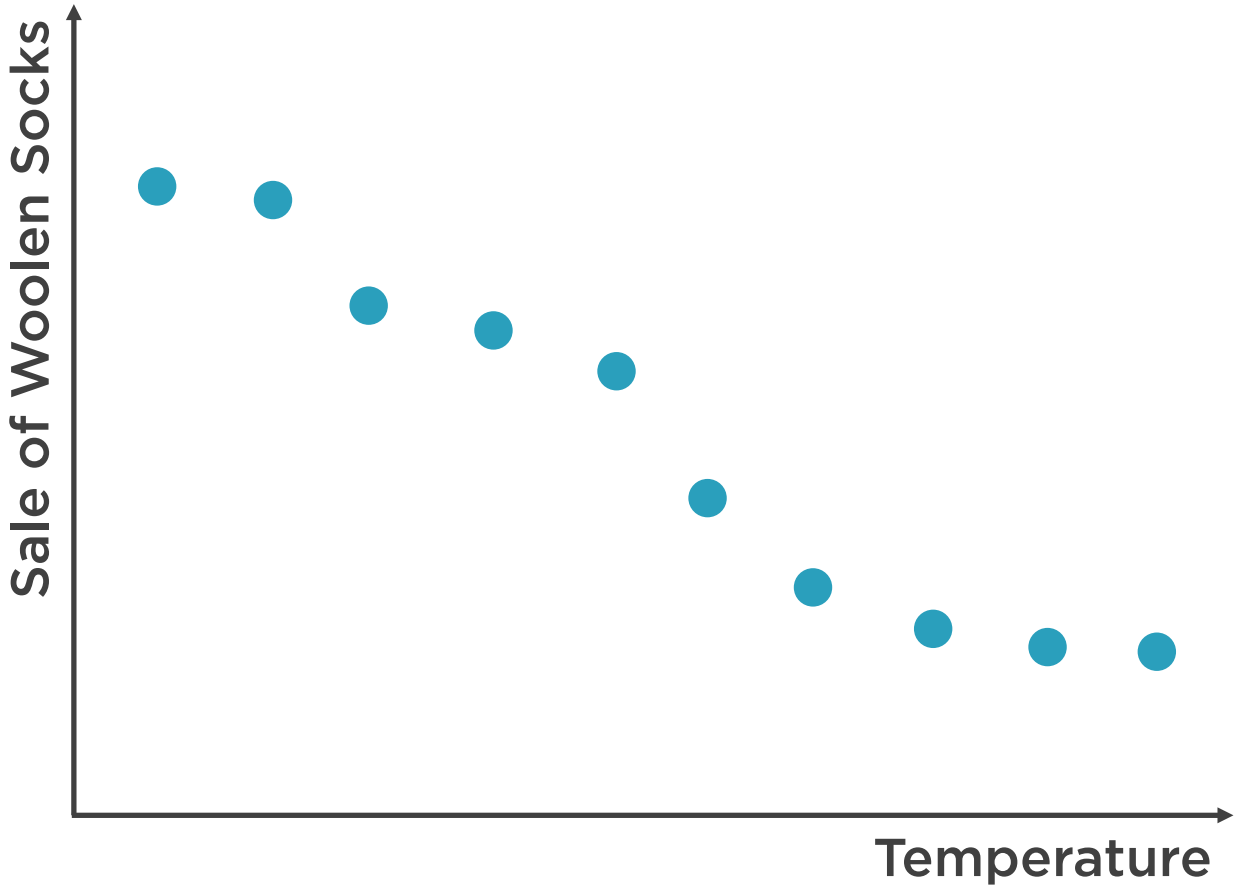


Train-test split produces high variations in test metrics

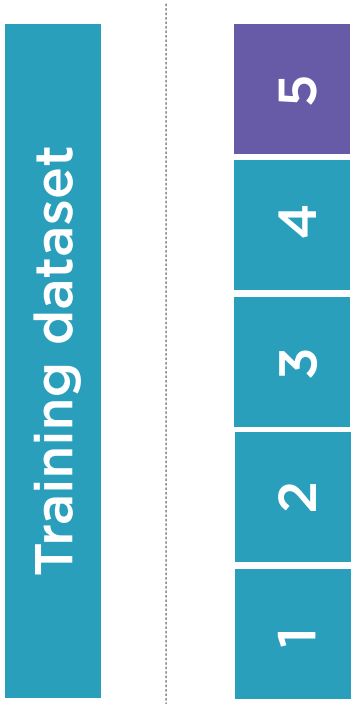
How Cross-validation Works?

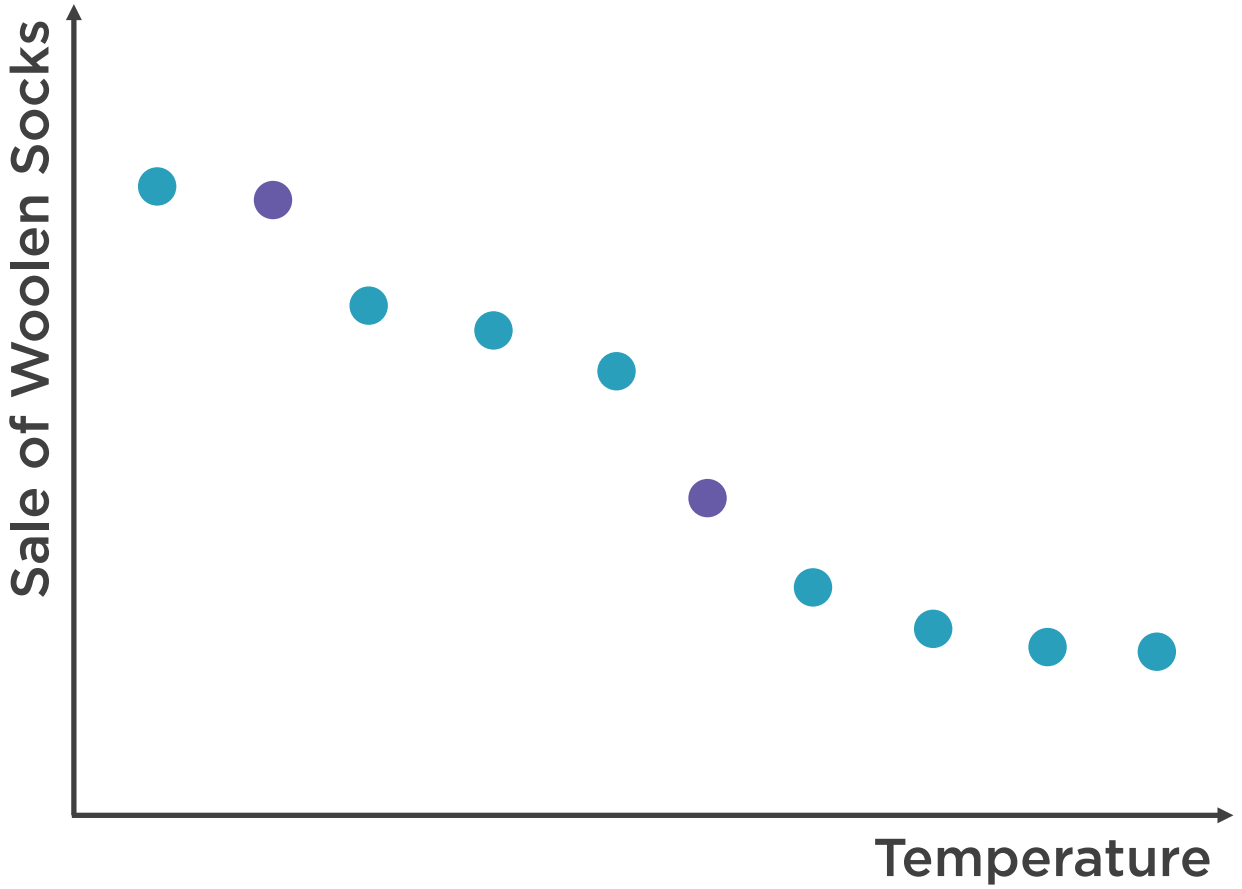


K = 5



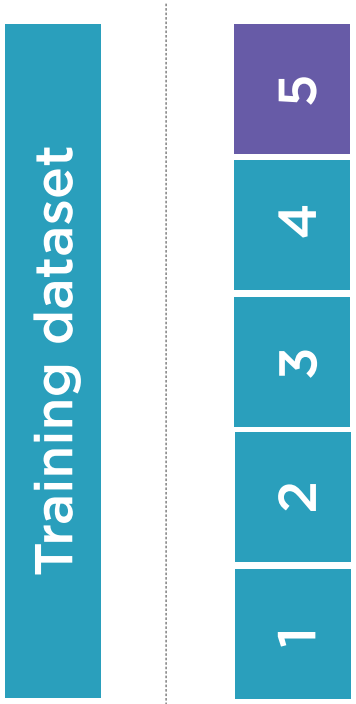
How Cross-validation Works?

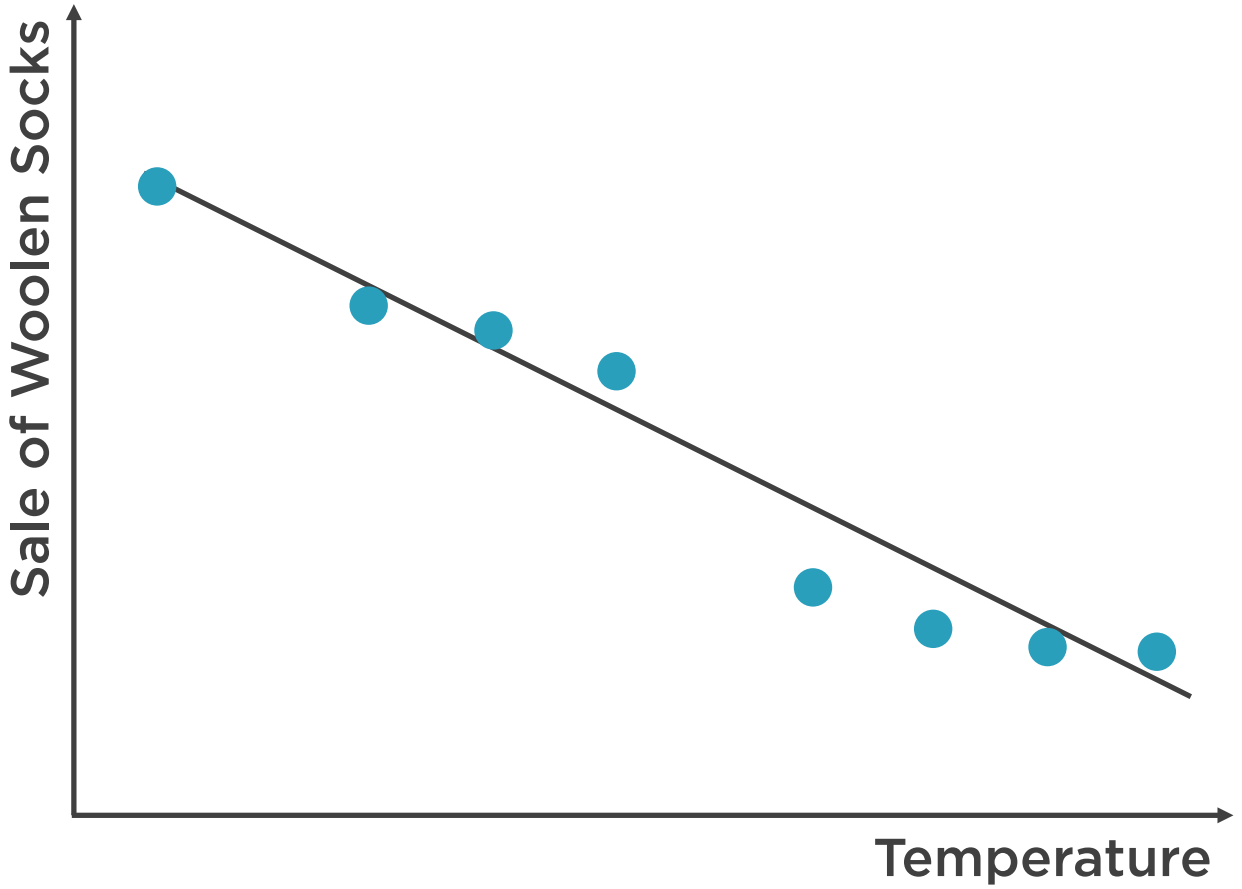




Training set Validation fold

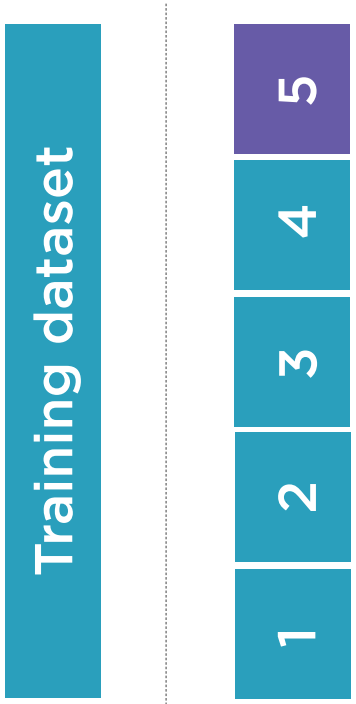
How Cross-validation Works?

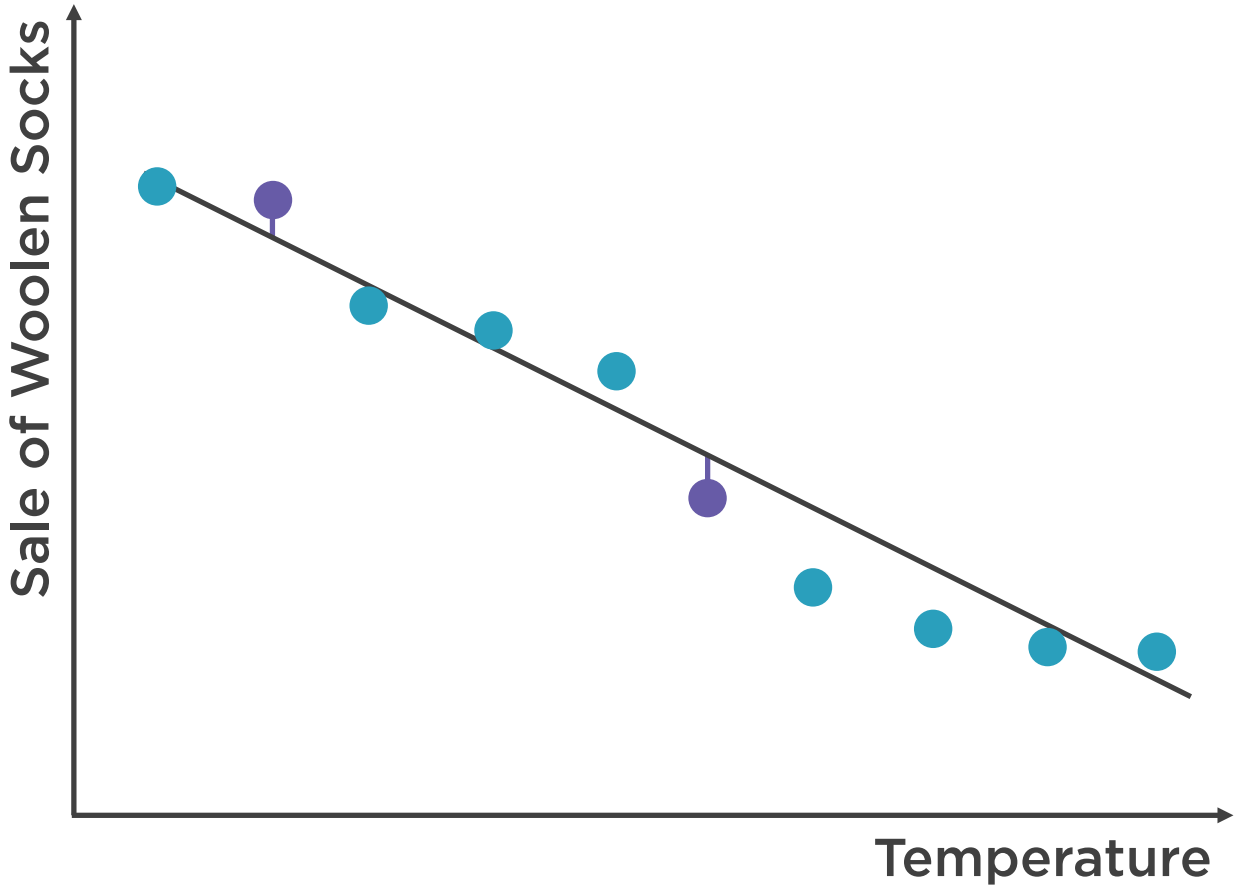




Training set Validation fold

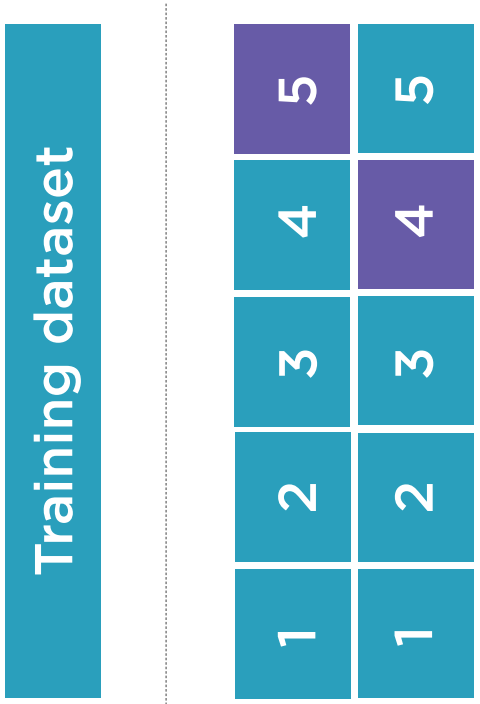
How Cross-validation Works?

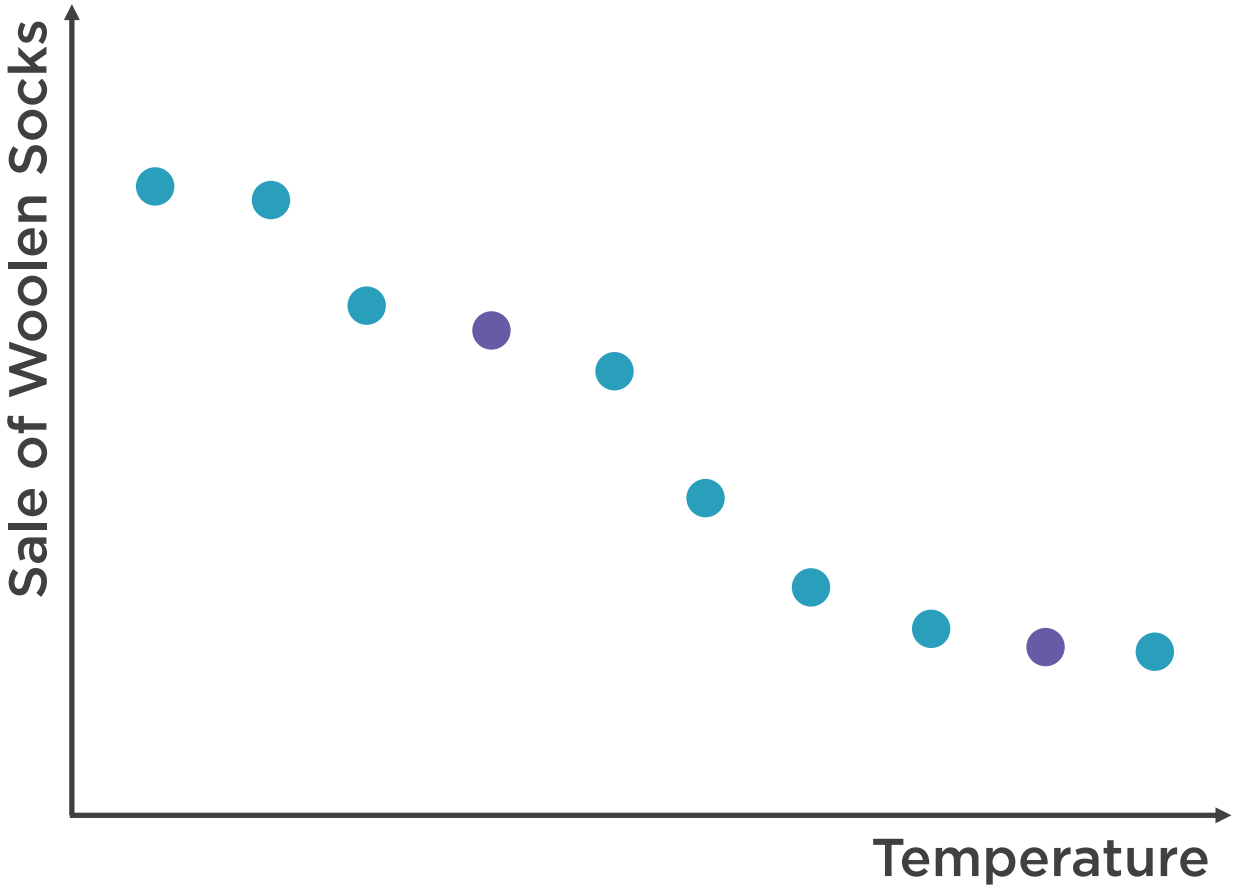




Training set Validation fold

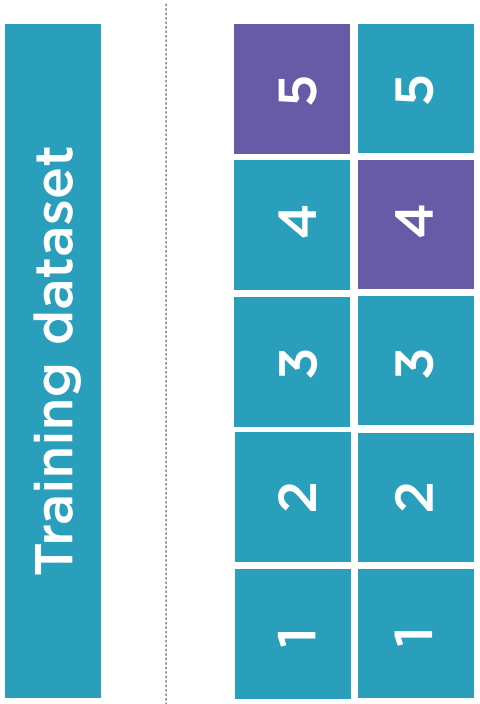
How Cross-validation Works?

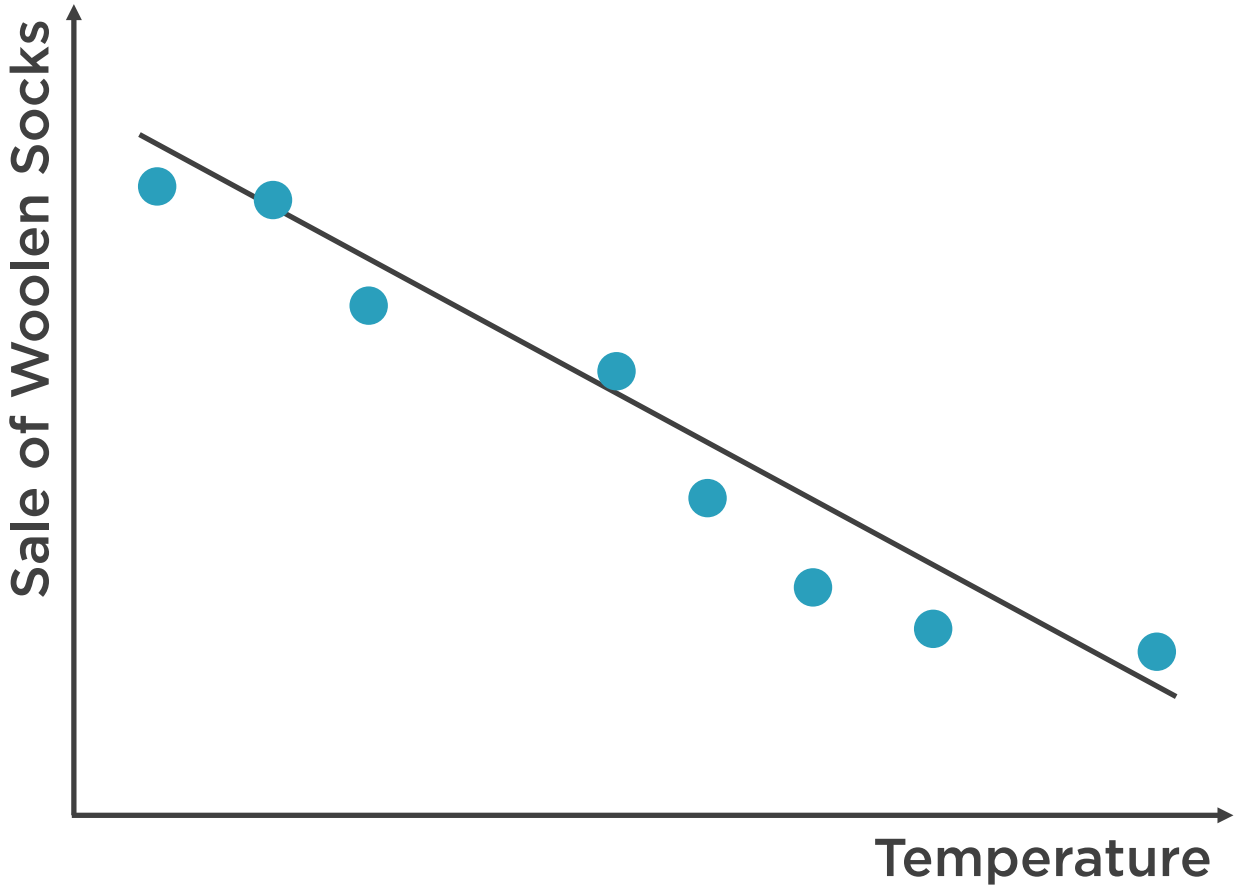




Training set Validation fold

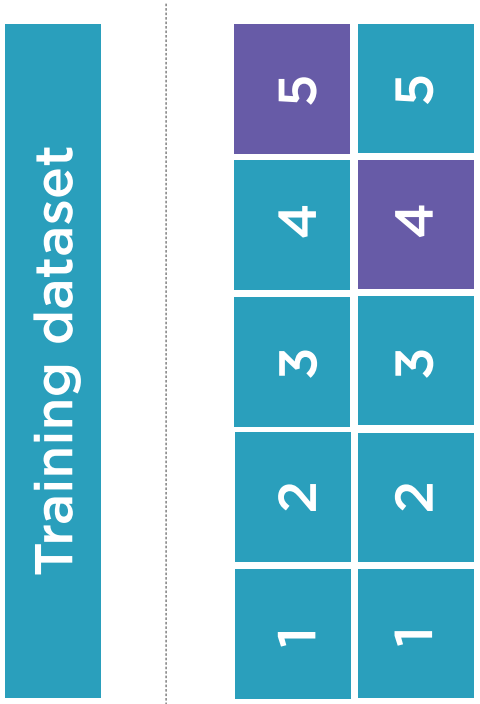
How Cross-validation Works?

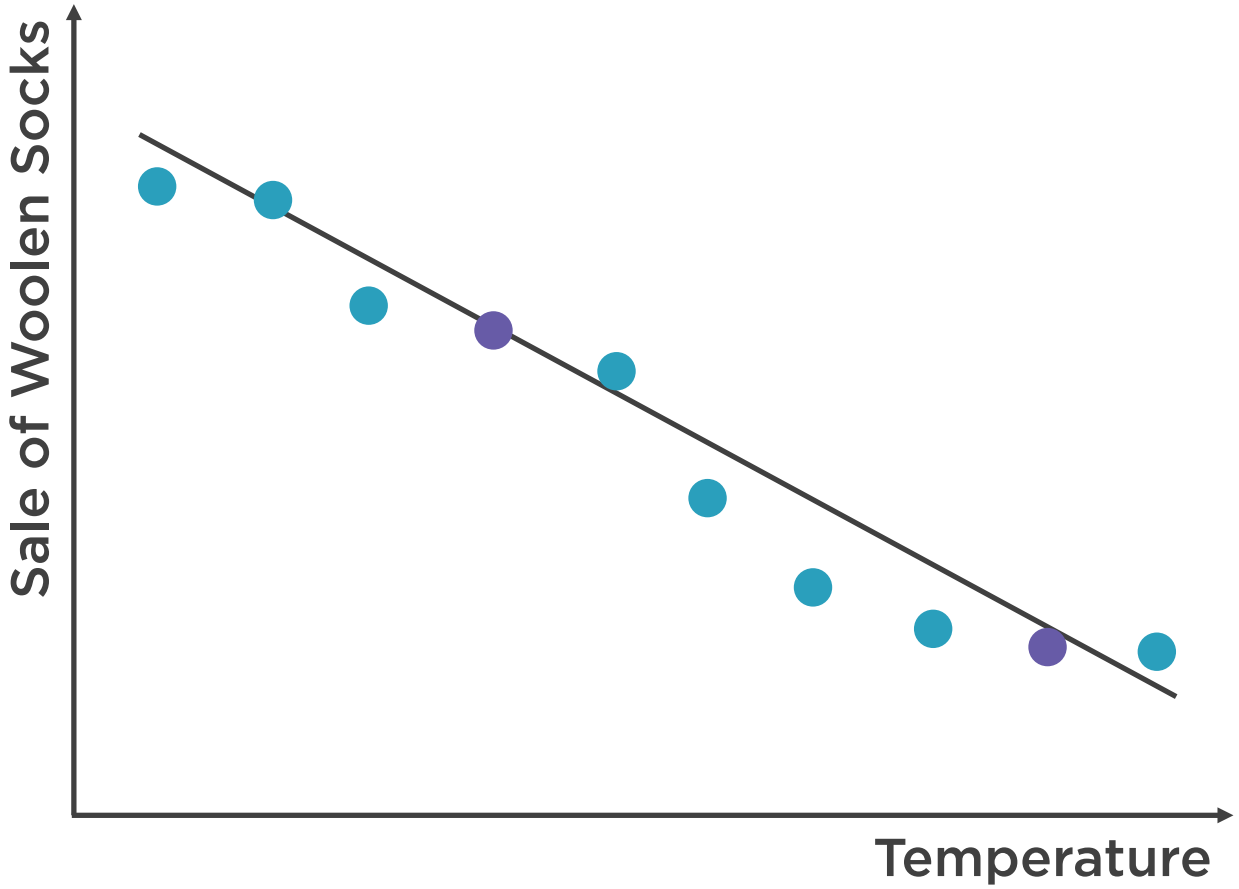




Training set Validation fold

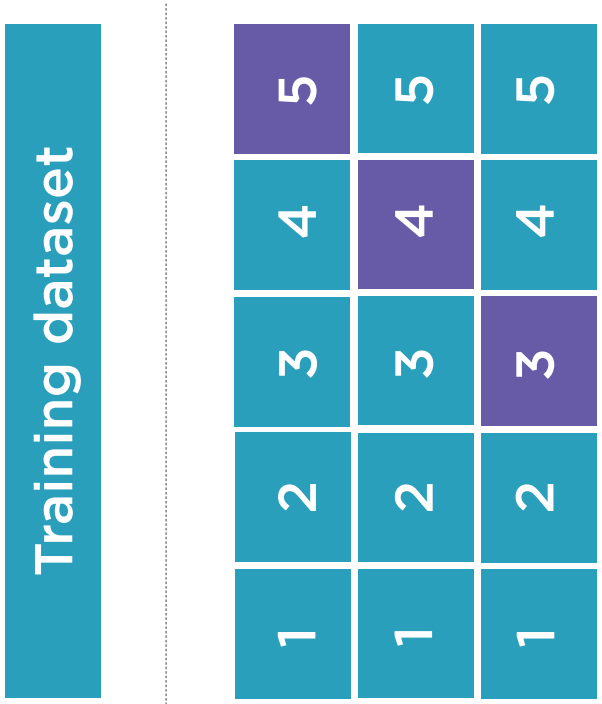
How Cross-validation Works?

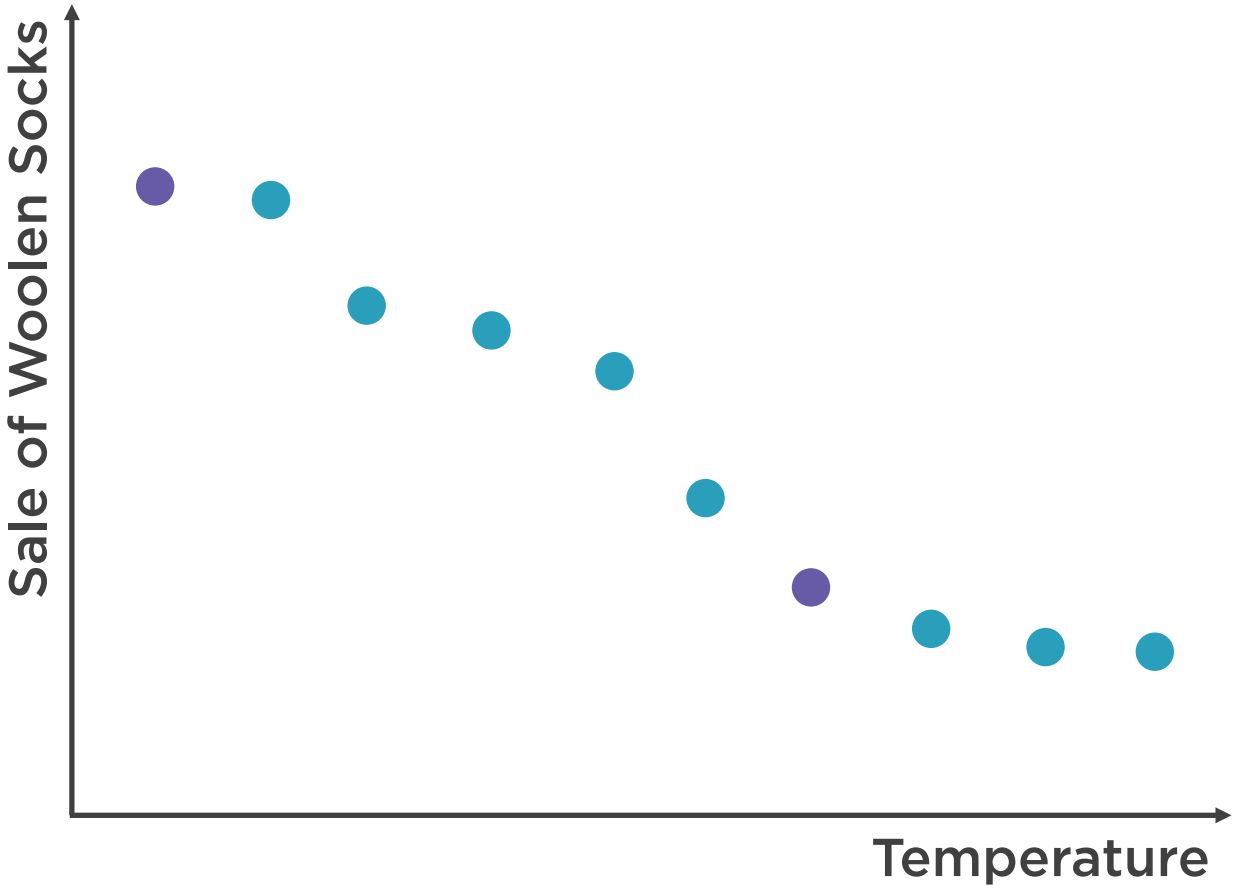




Training set Validation fold

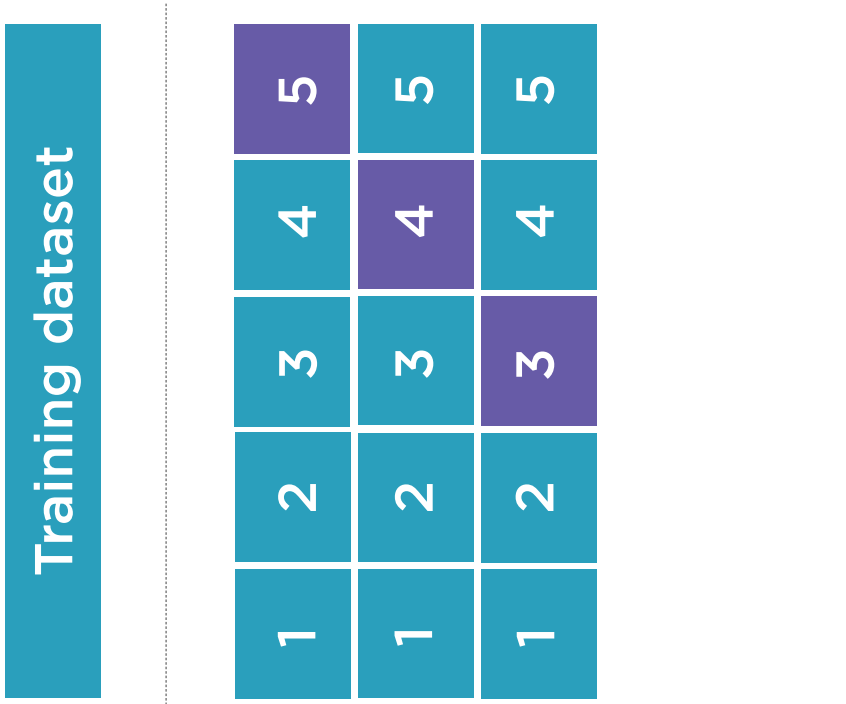
How Cross-validation Works?

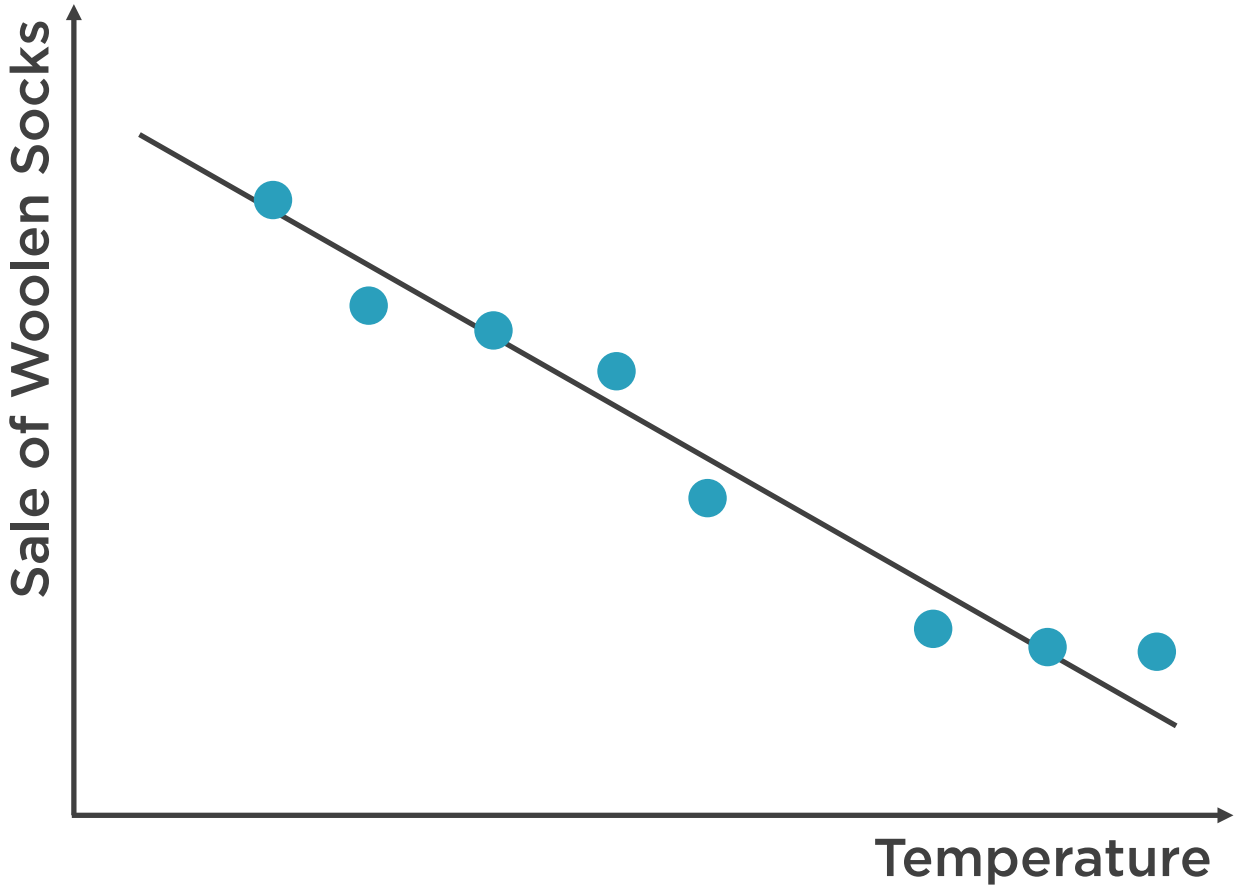




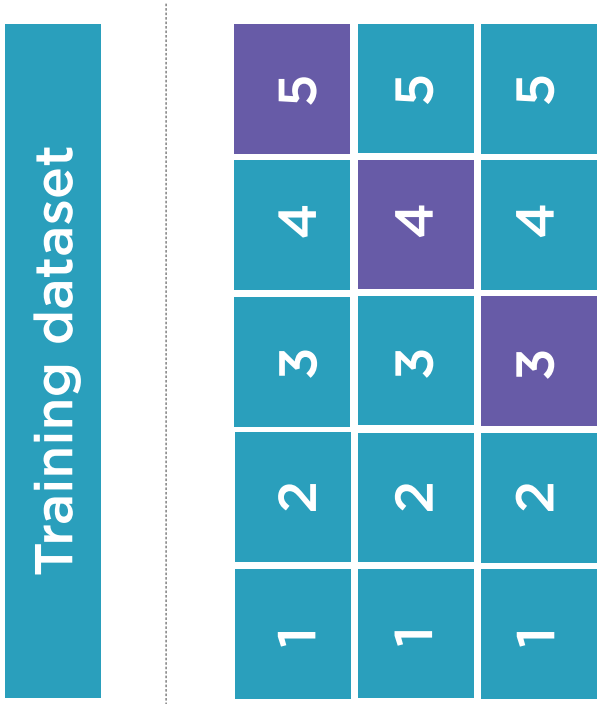
Training set Validation fold

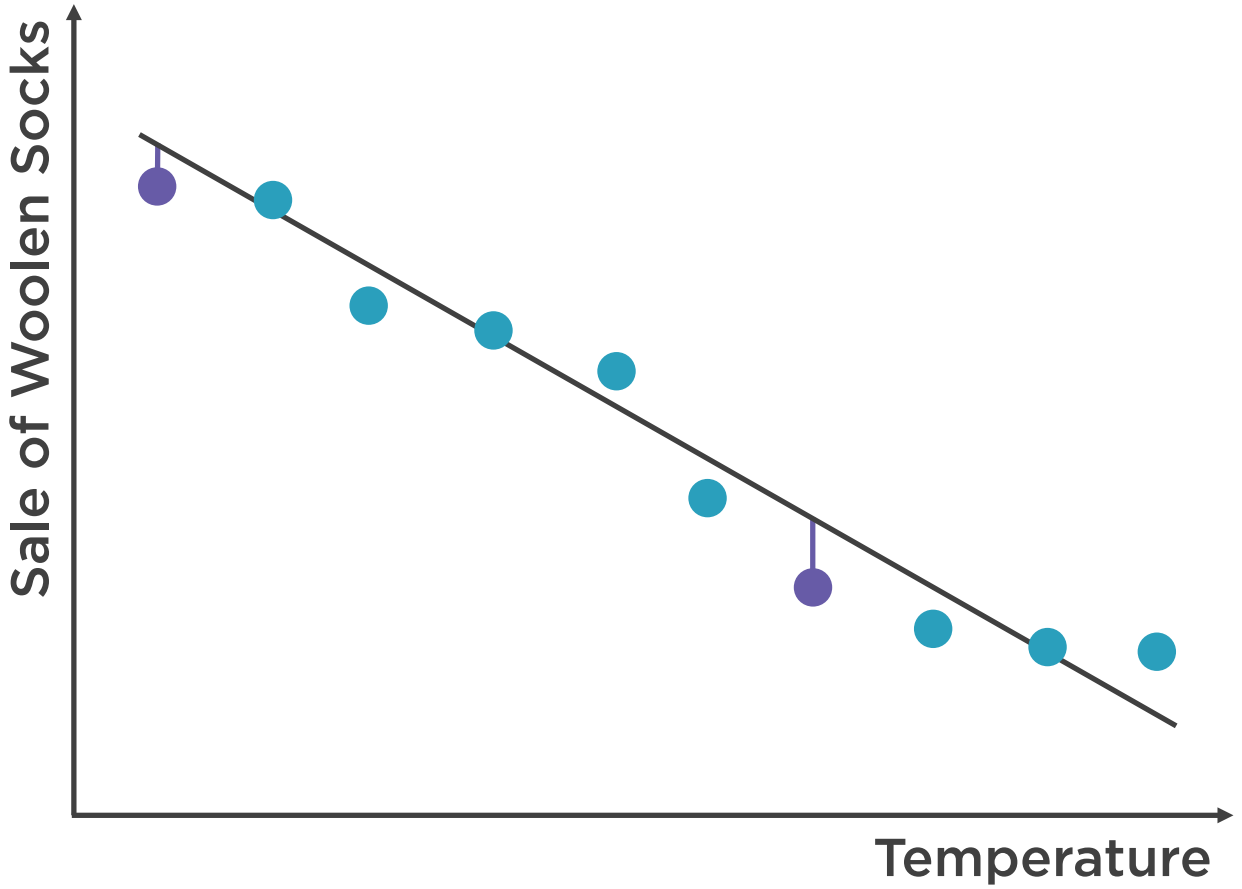
How Cross-validation Works?





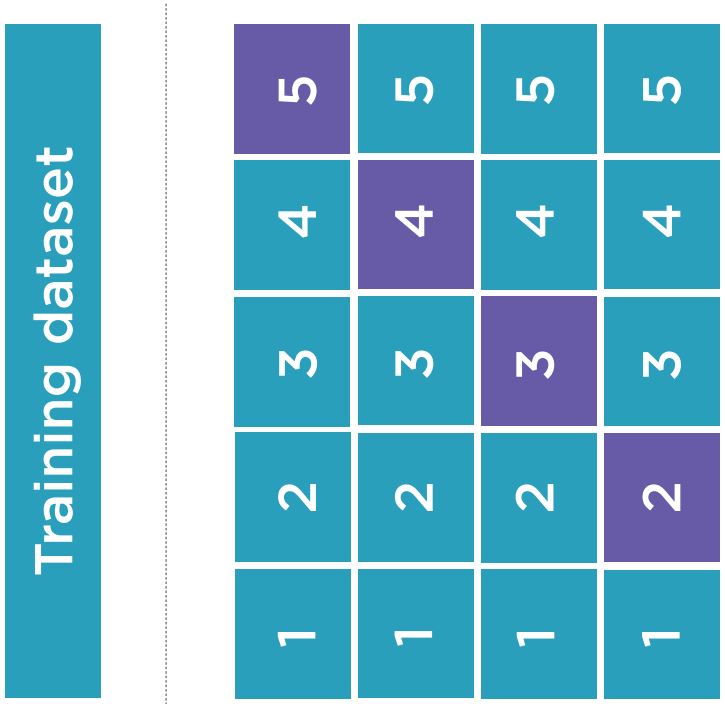
How Cross-validation Works?

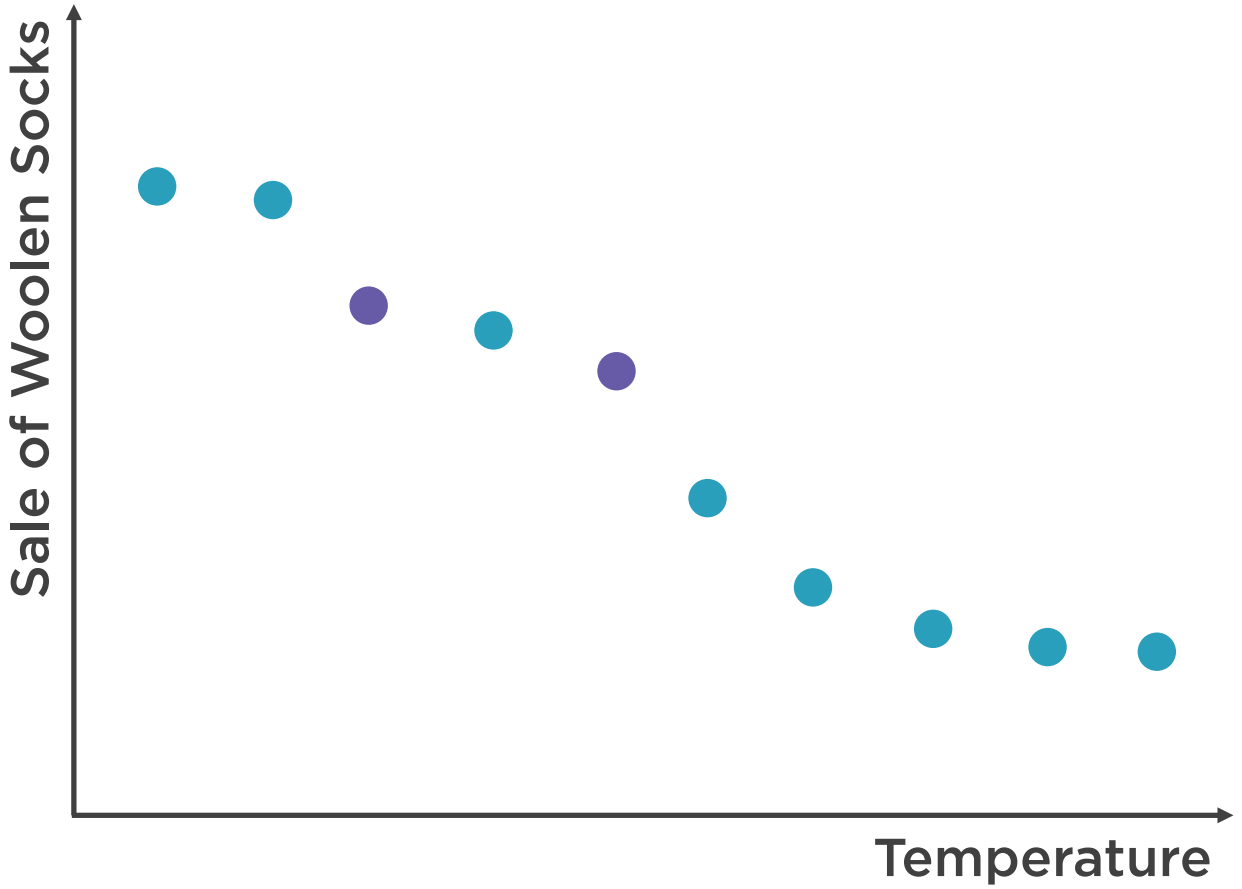




Training set Validation fold

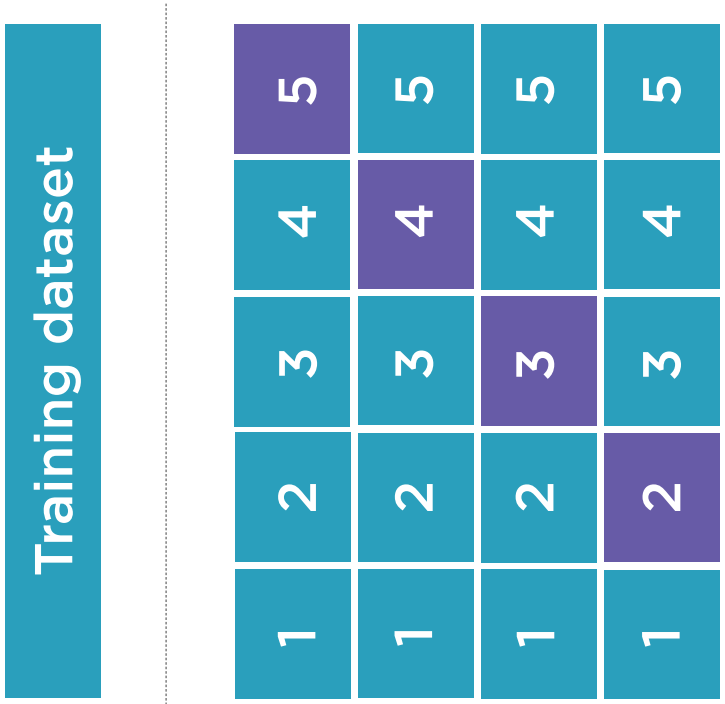
How Cross-validation Works?

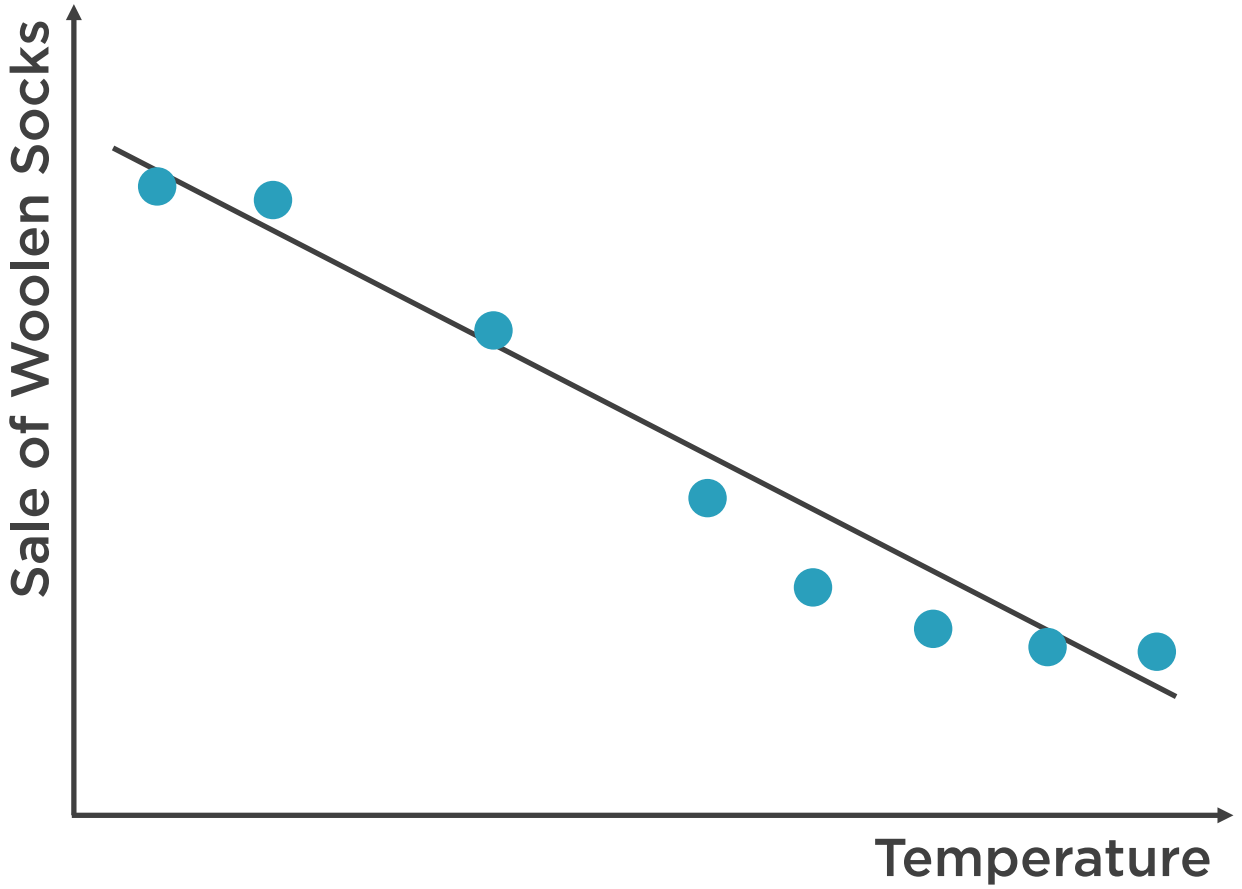




Training set Validation fold

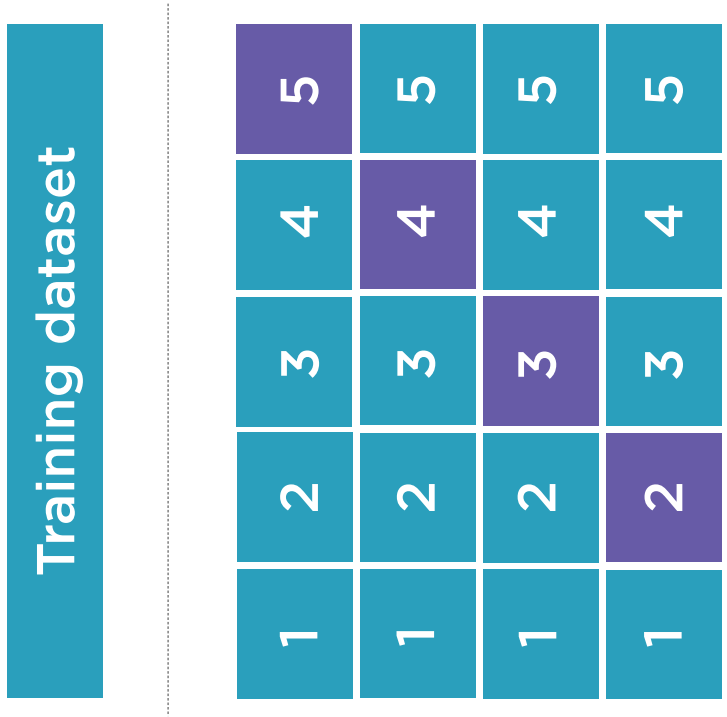
How Cross-validation Works?

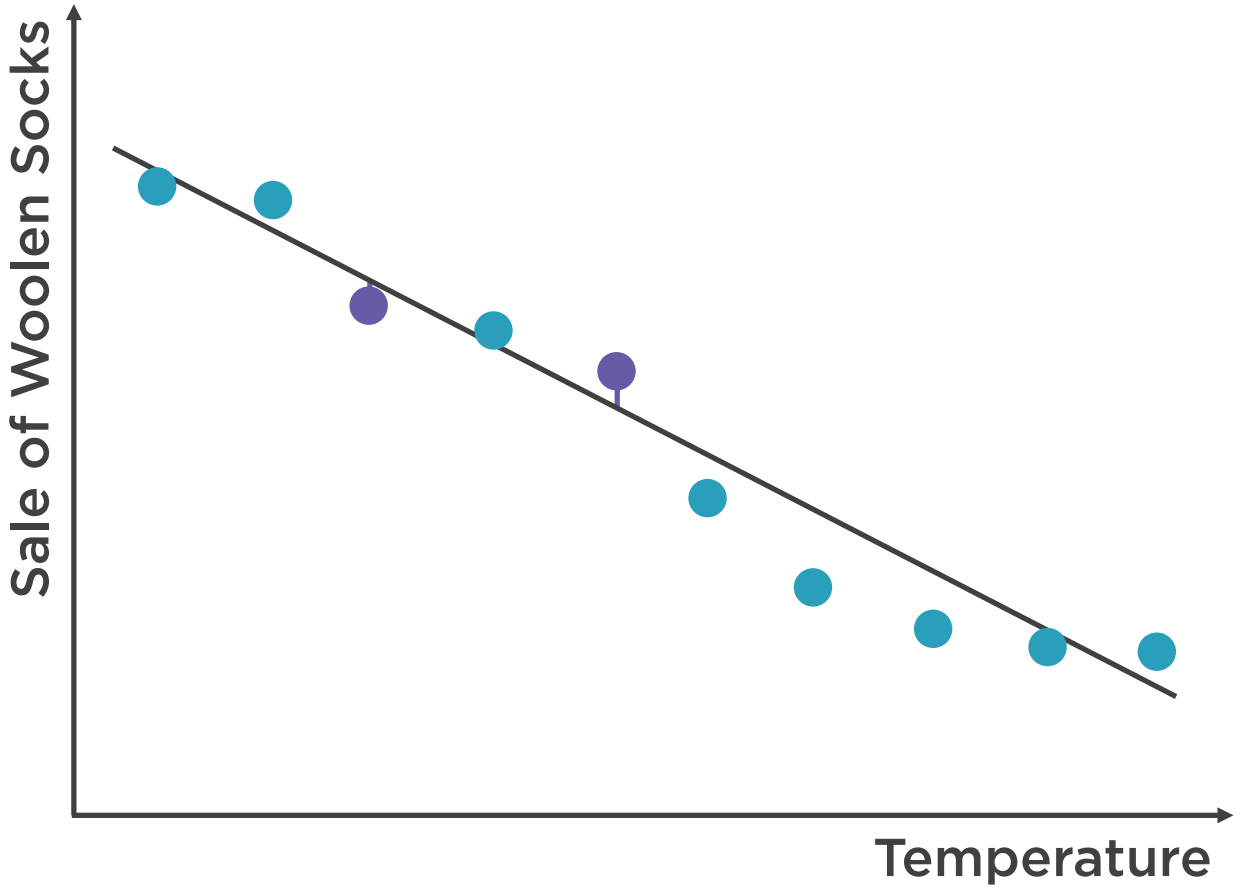




Training set Validation fold

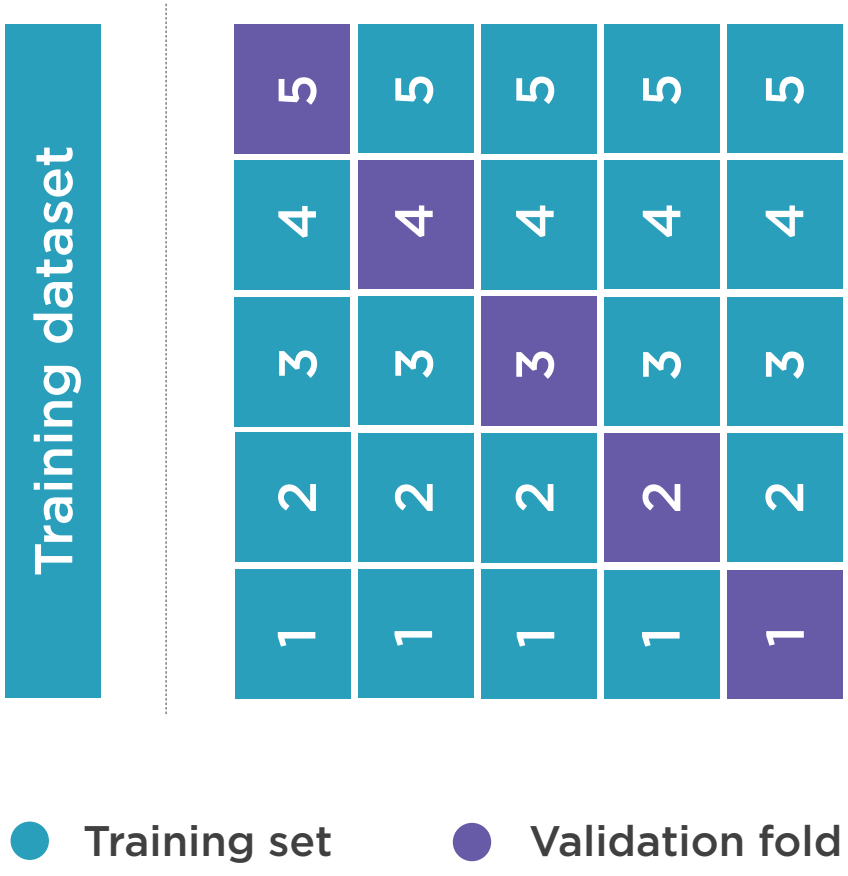
How Cross-validation Works?

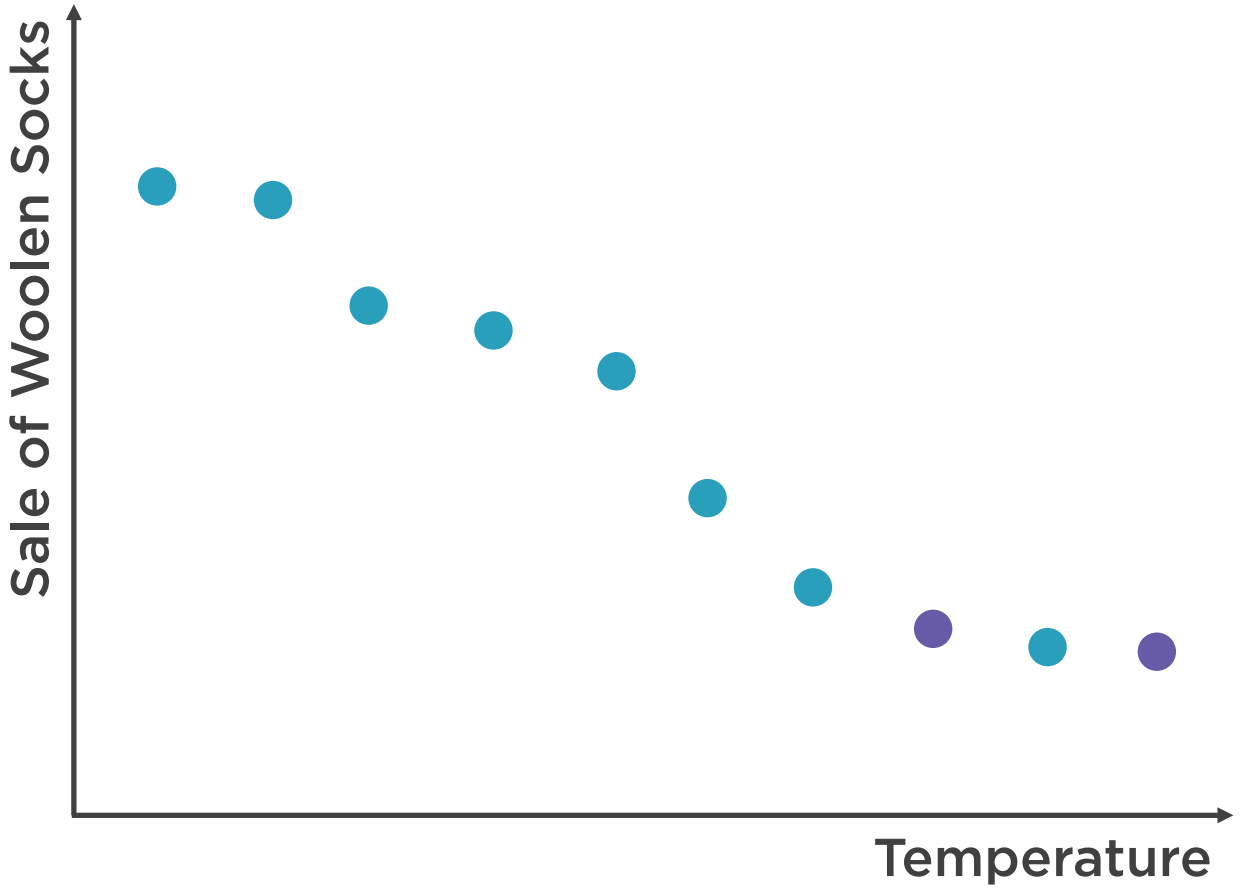




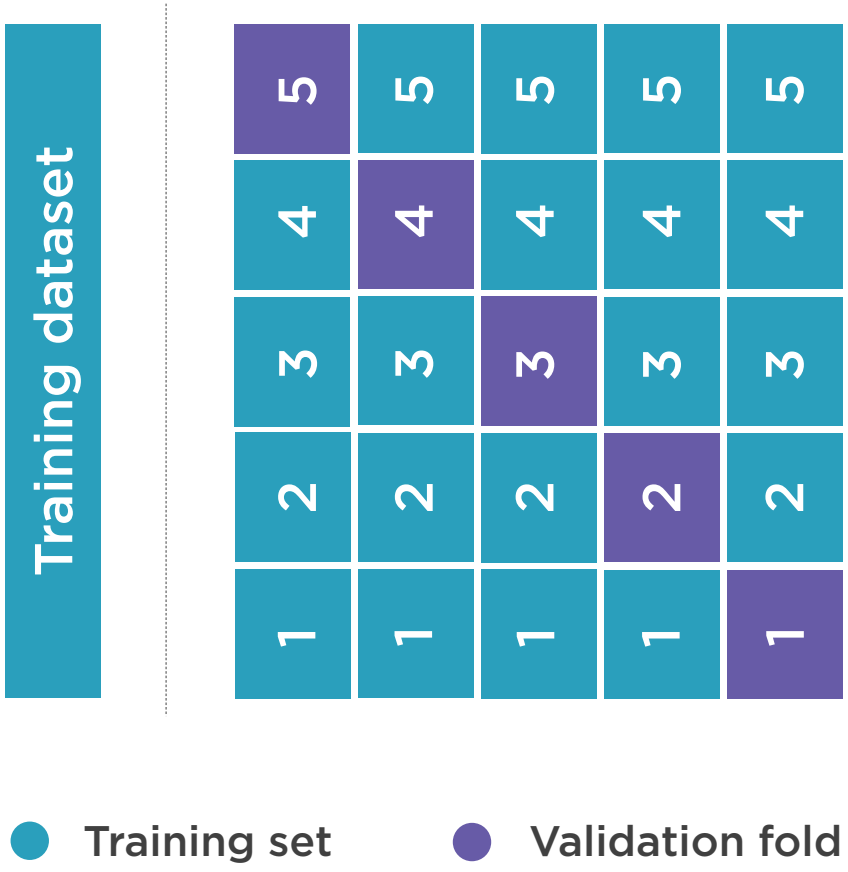
Training set Validation fold

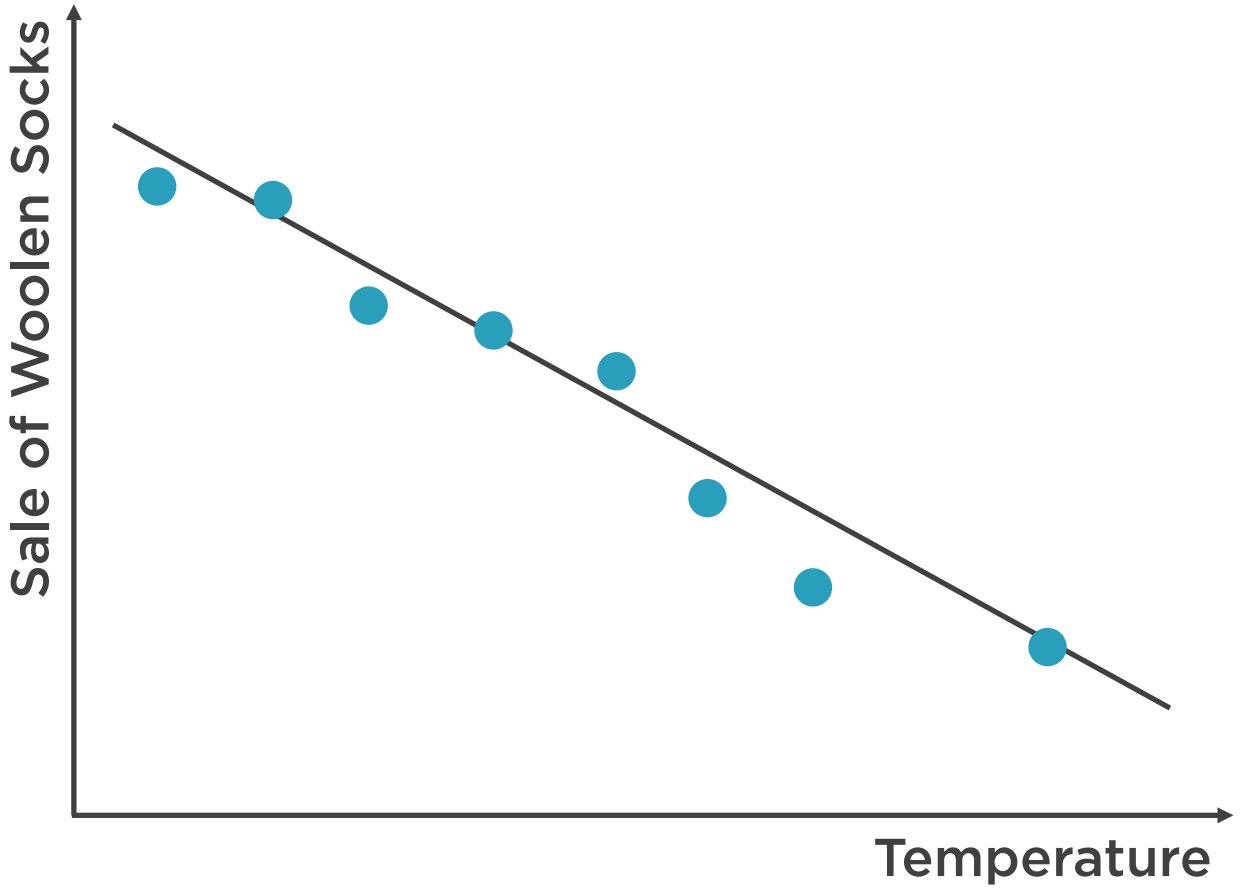
How Cross-validation Works?



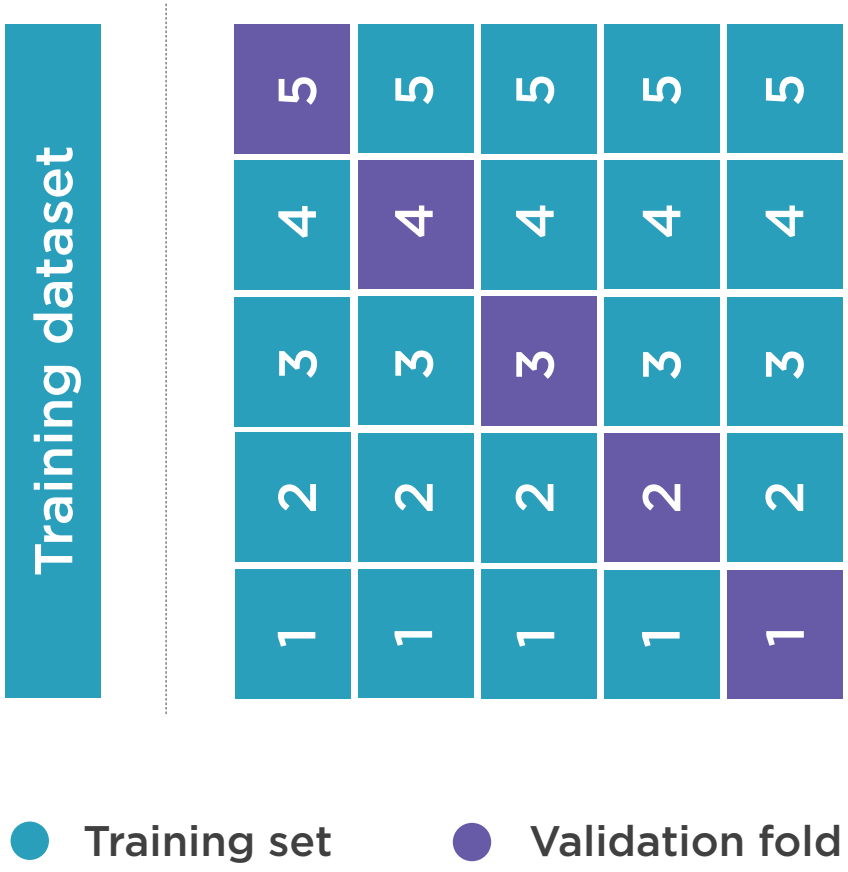


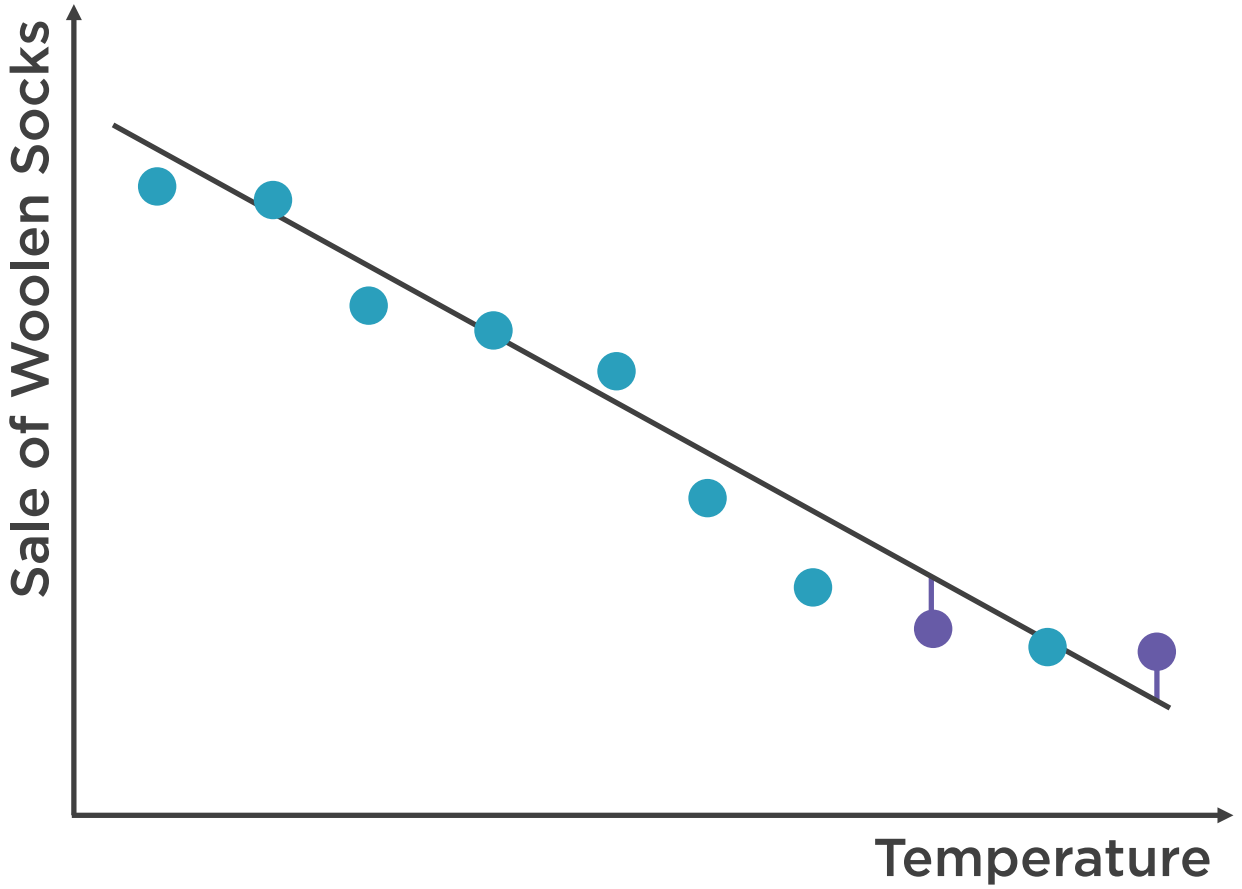
How Cross-validation Works?



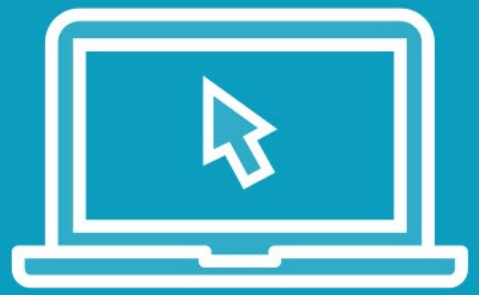


How Cross-validation Works?





Demo

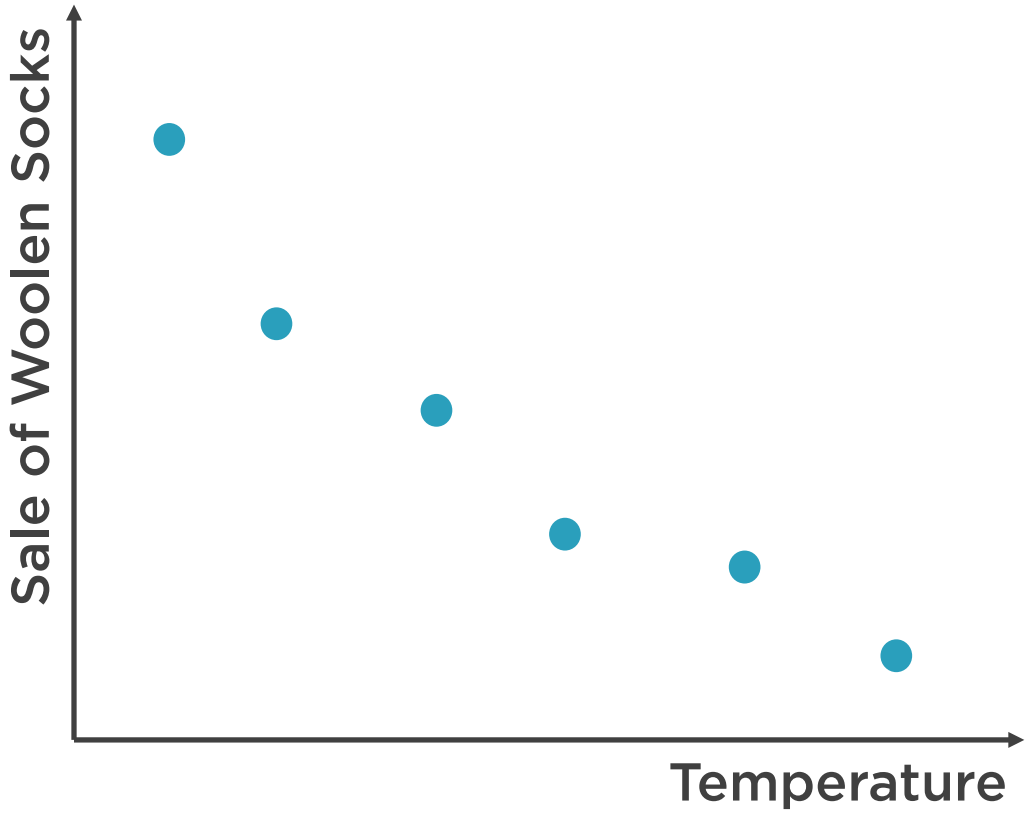


Model Selection – Evaluate 2 models

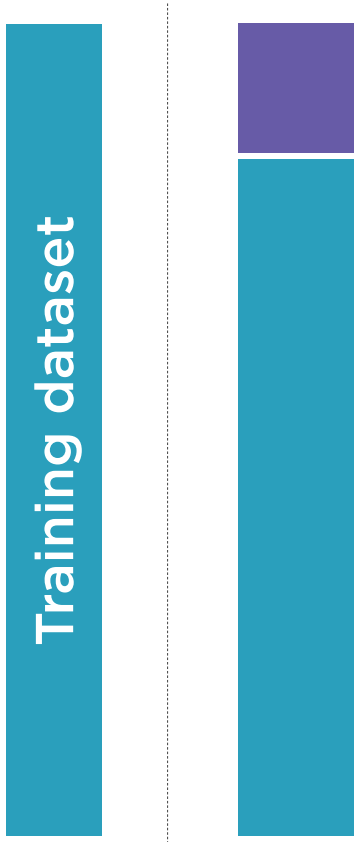
Leave-one-out Cross Validation

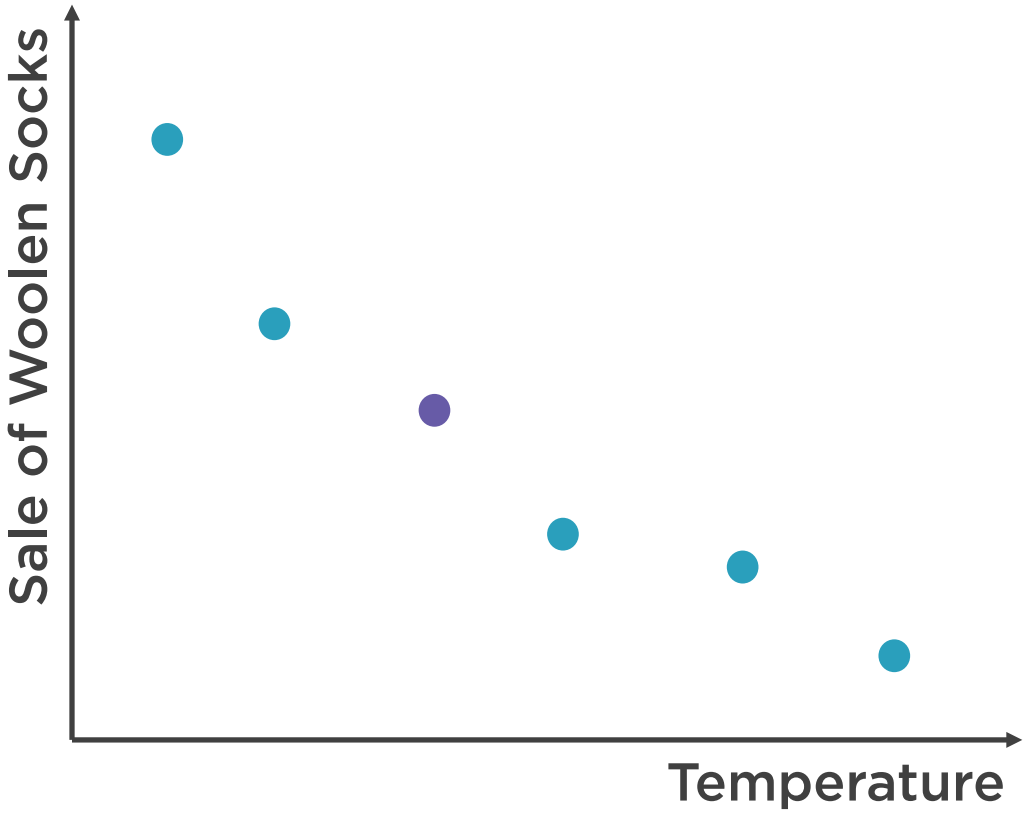
Leave-one-out Cross Validation (LOOCV)

Training dataset



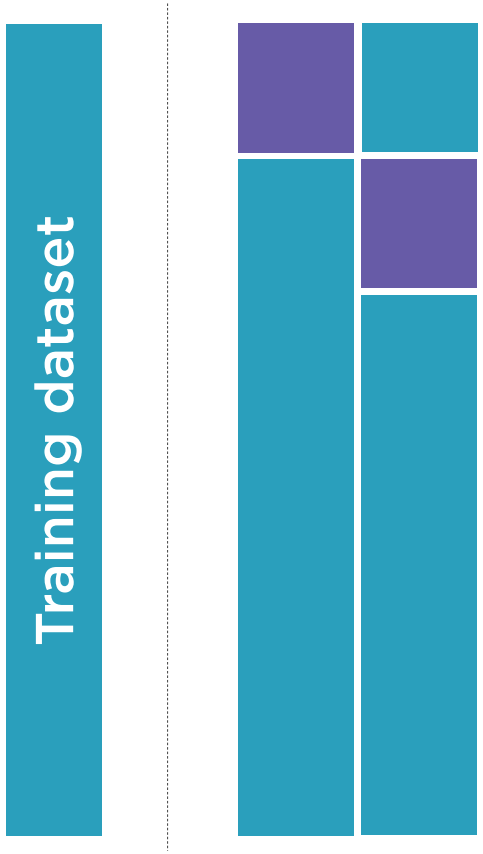
Leave-one-out Cross Validation (LOOCV)

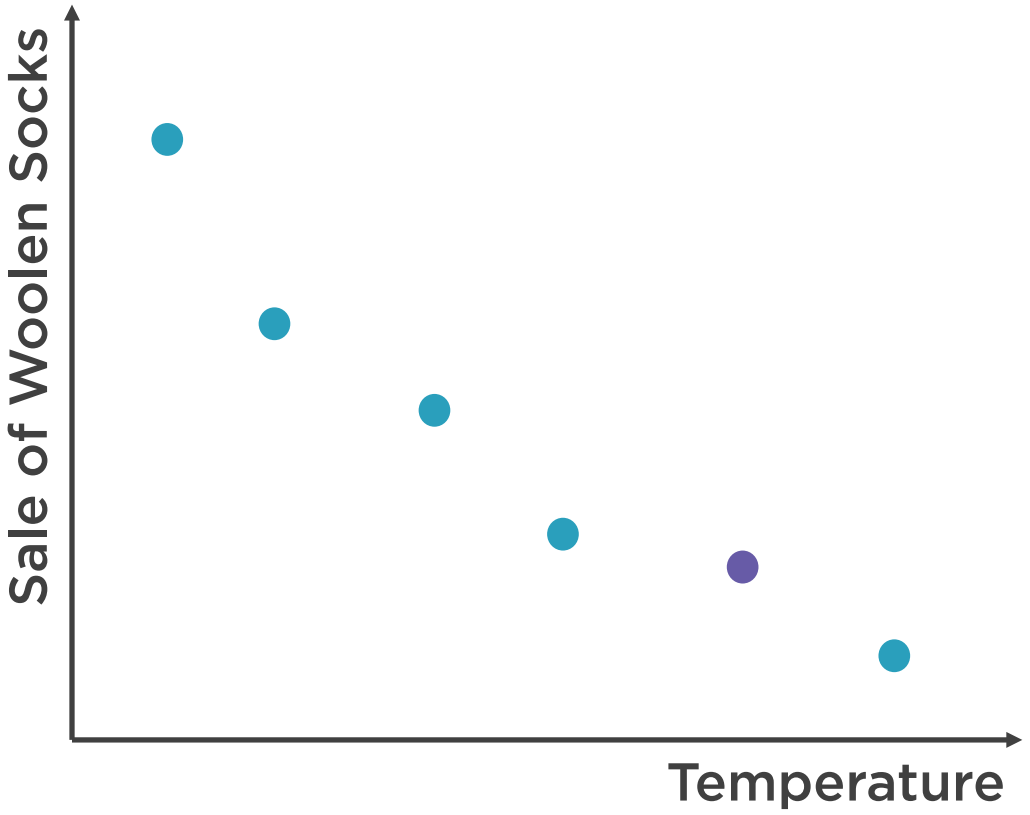




Training set Validation data point

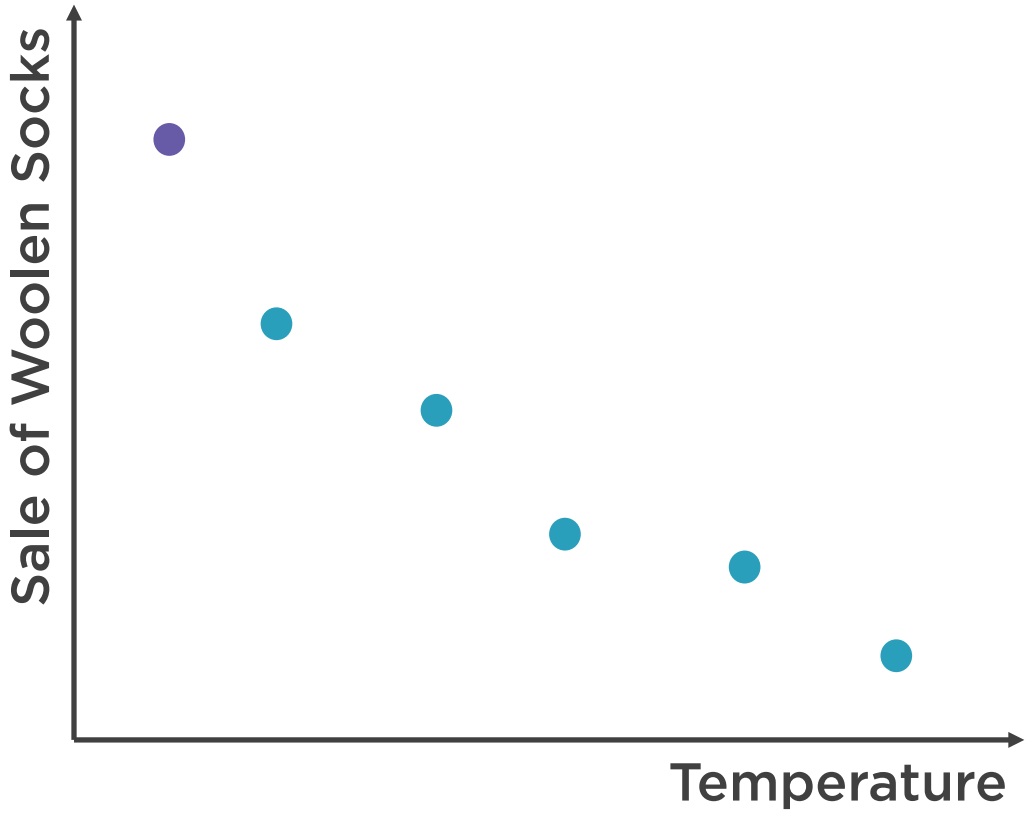
Leave-one-out Cross Validation (LOOCV)





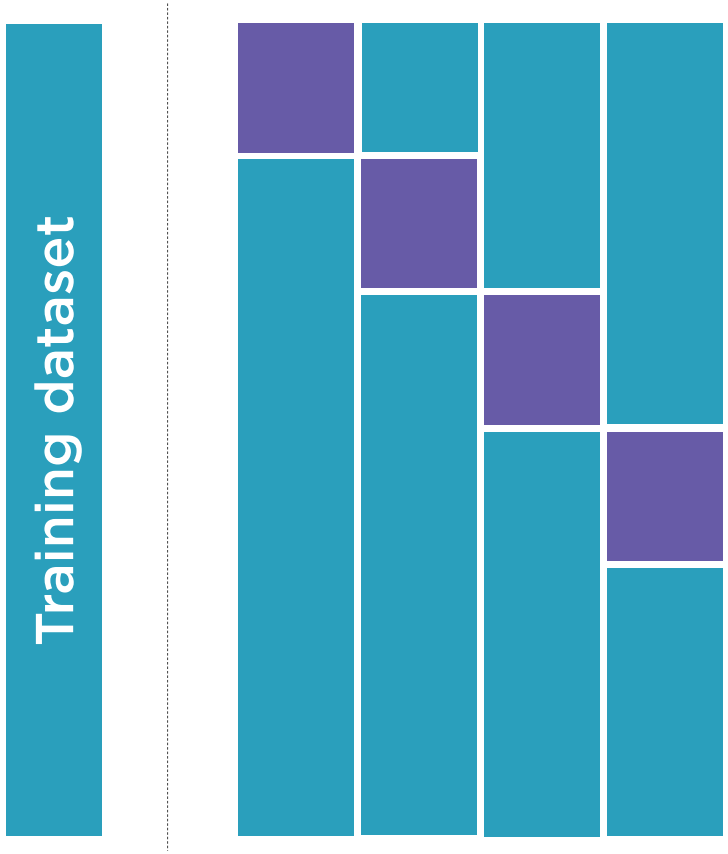
Training set Validation data point

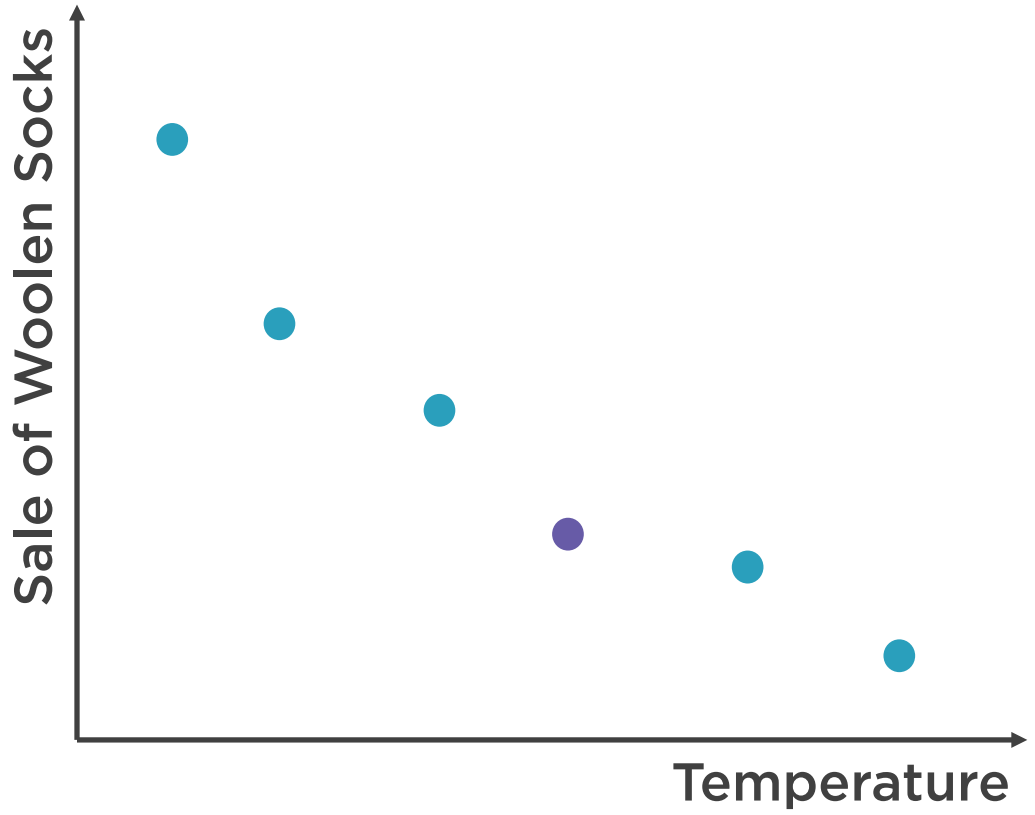
Leave-one-out Cross Validation (LOOCV)



Training set Validation data point

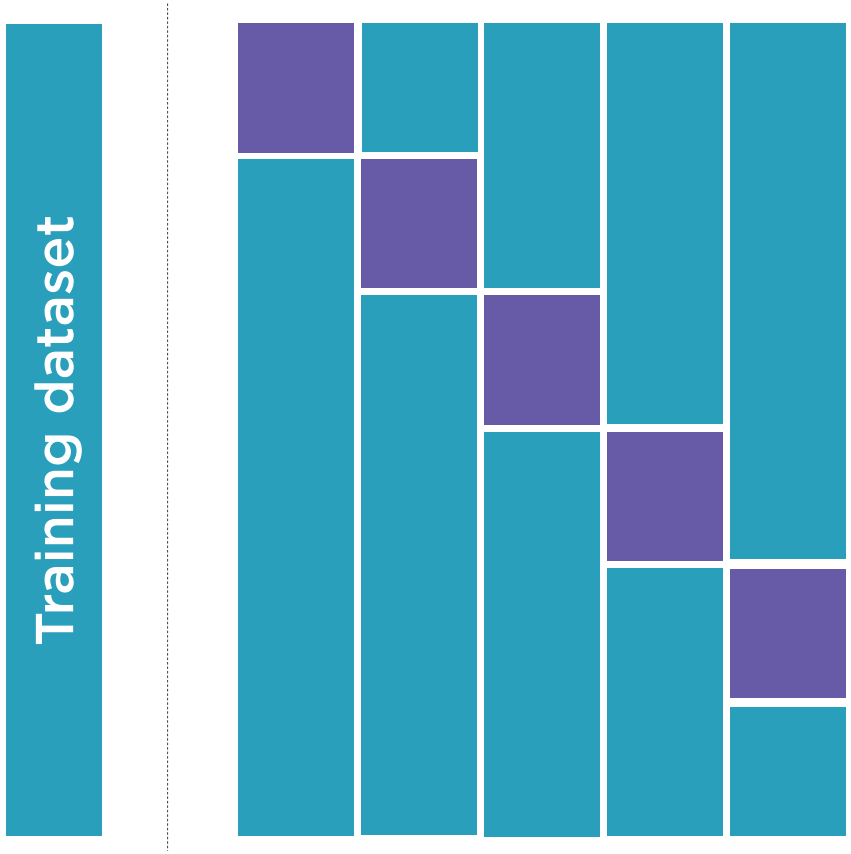
Leave-one-out Cross Validation (LOOCV)

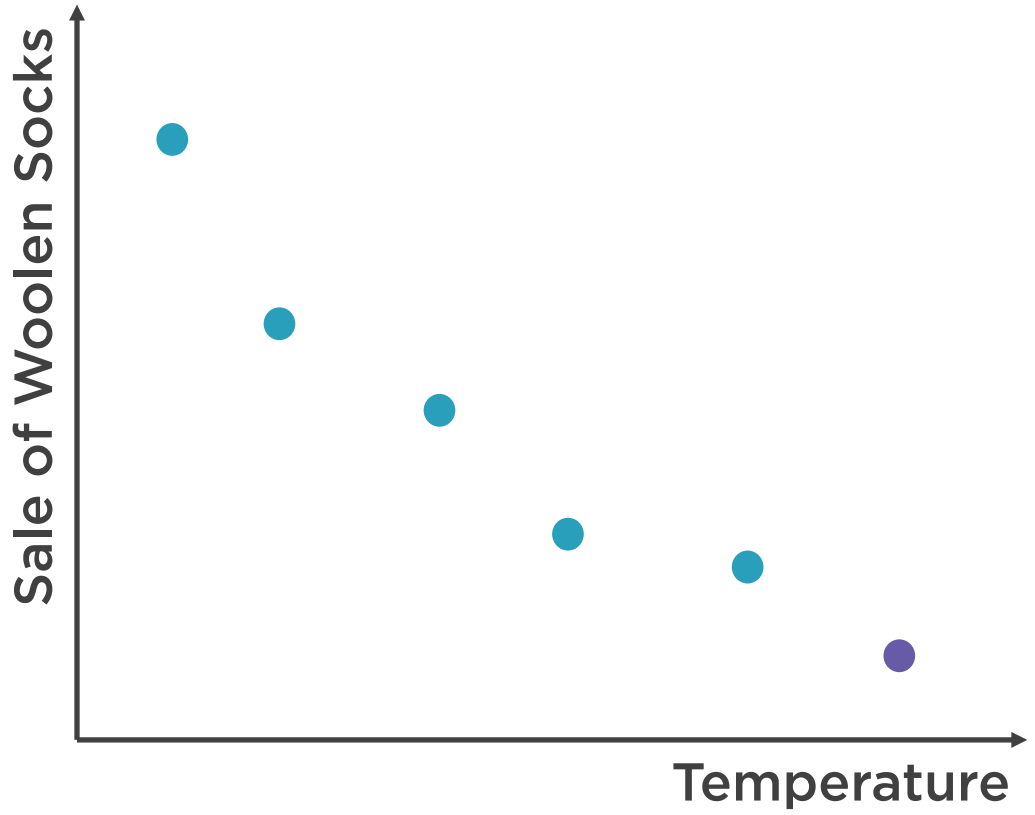




Training set Validation data point

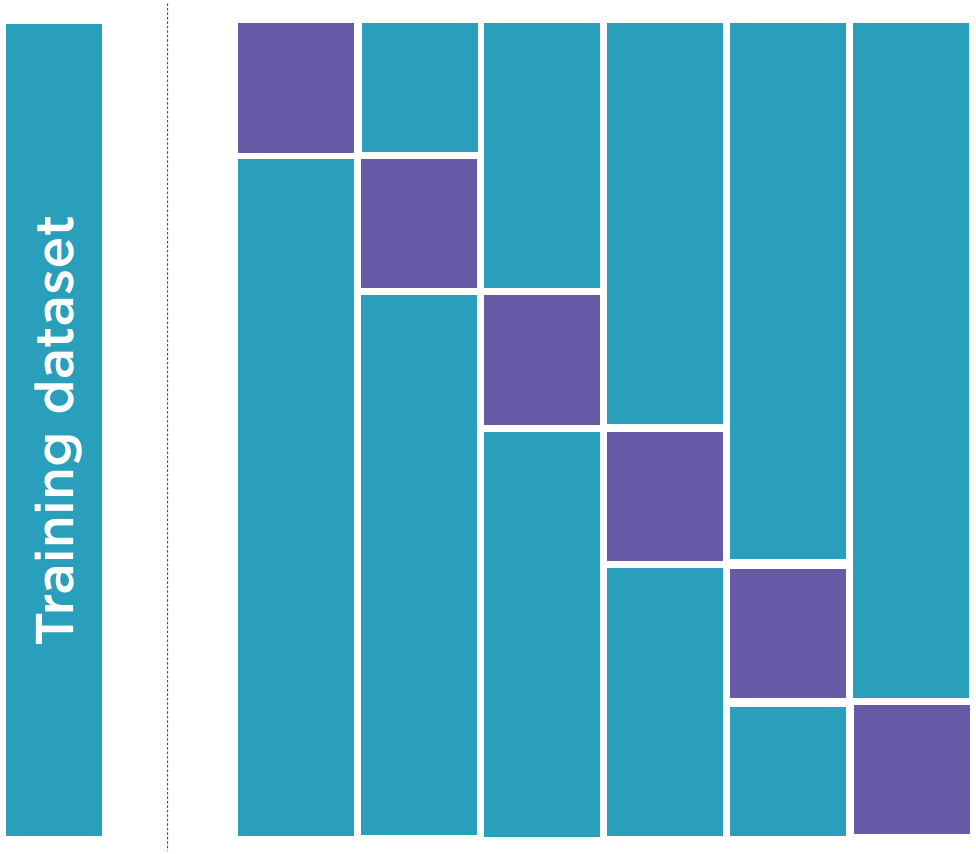
Leave-one-out Cross Validation (LOOCV)

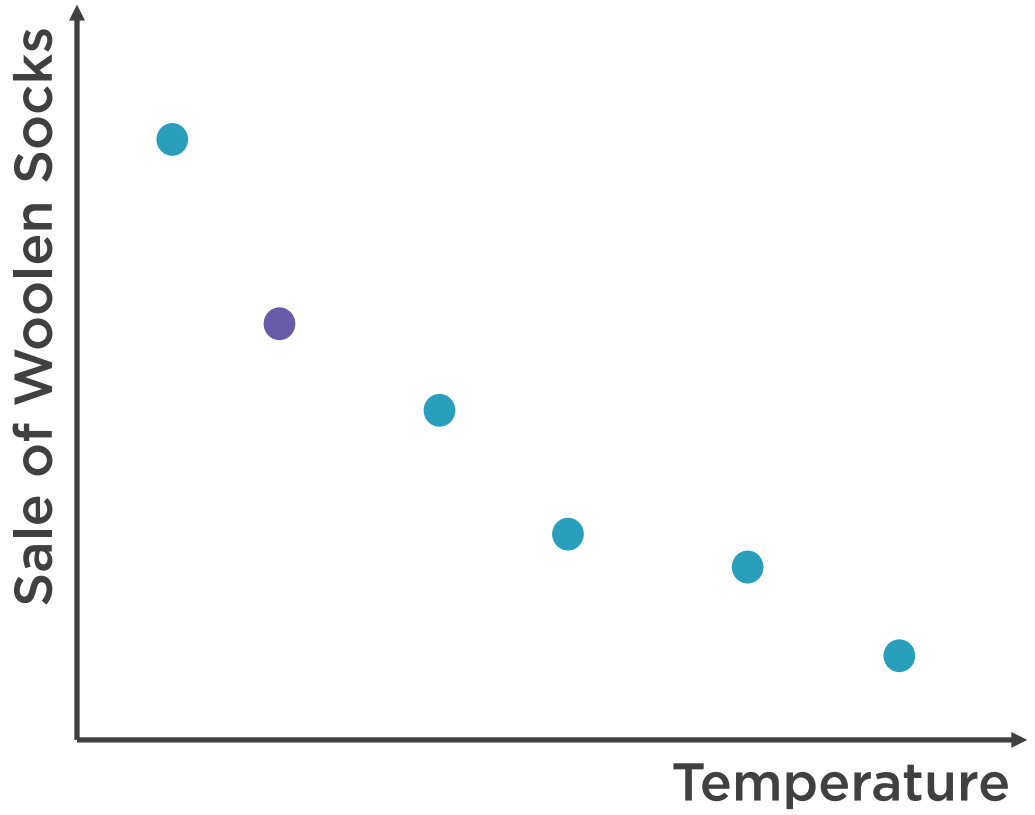




Training set Validation data point

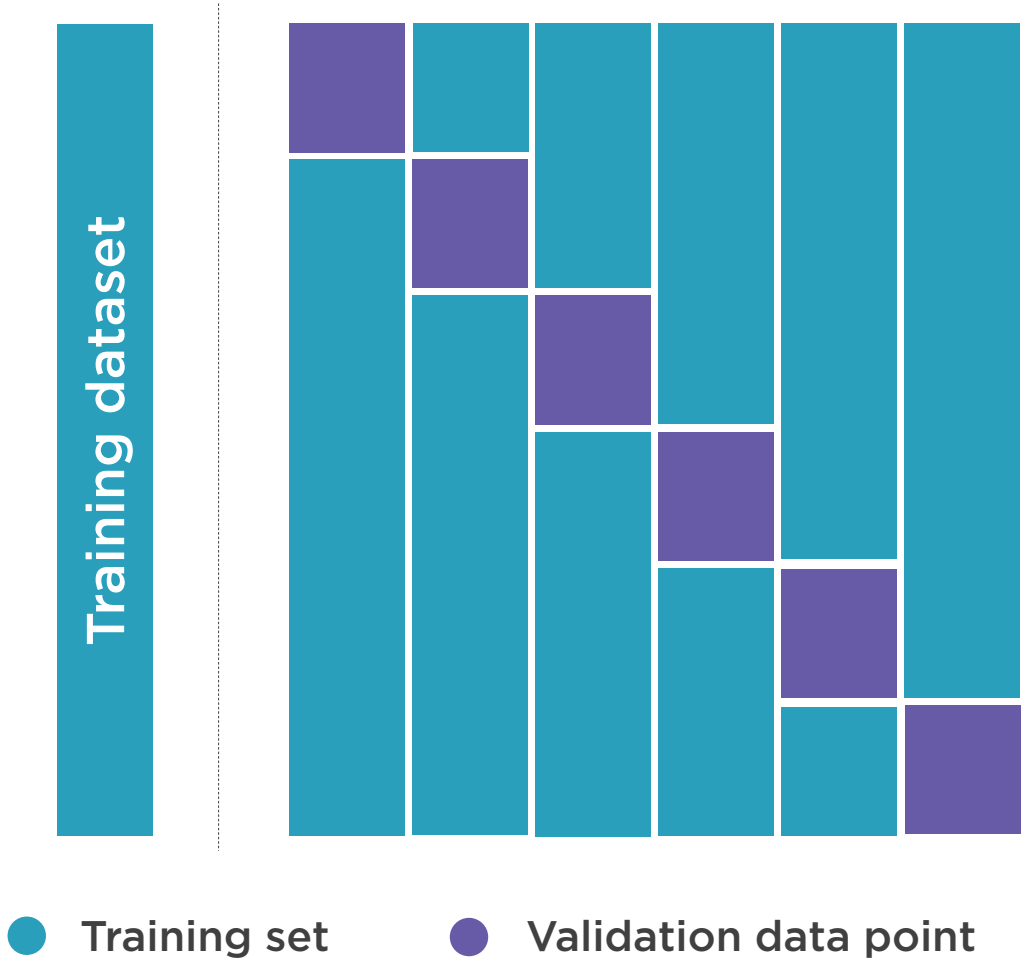
Leave-one-out Cross Validation (LOOCV)





Training set Validation data point

Leave-one-out Cross Validation (LOOCV)



- Less Bias
- High variance test error
- Use LOOCV for small datasets

Summary

Train -test split is simple & easy to use

Cross-validation efficiently uses the data

Randomize the data splits

Model selection to decide the best model

Remove duplicate records before splitting