GBUS 738 Data Mining

Discriminant Analysis and KNN

David Svancer – George Mason University School of Business

George Mason University School of Business

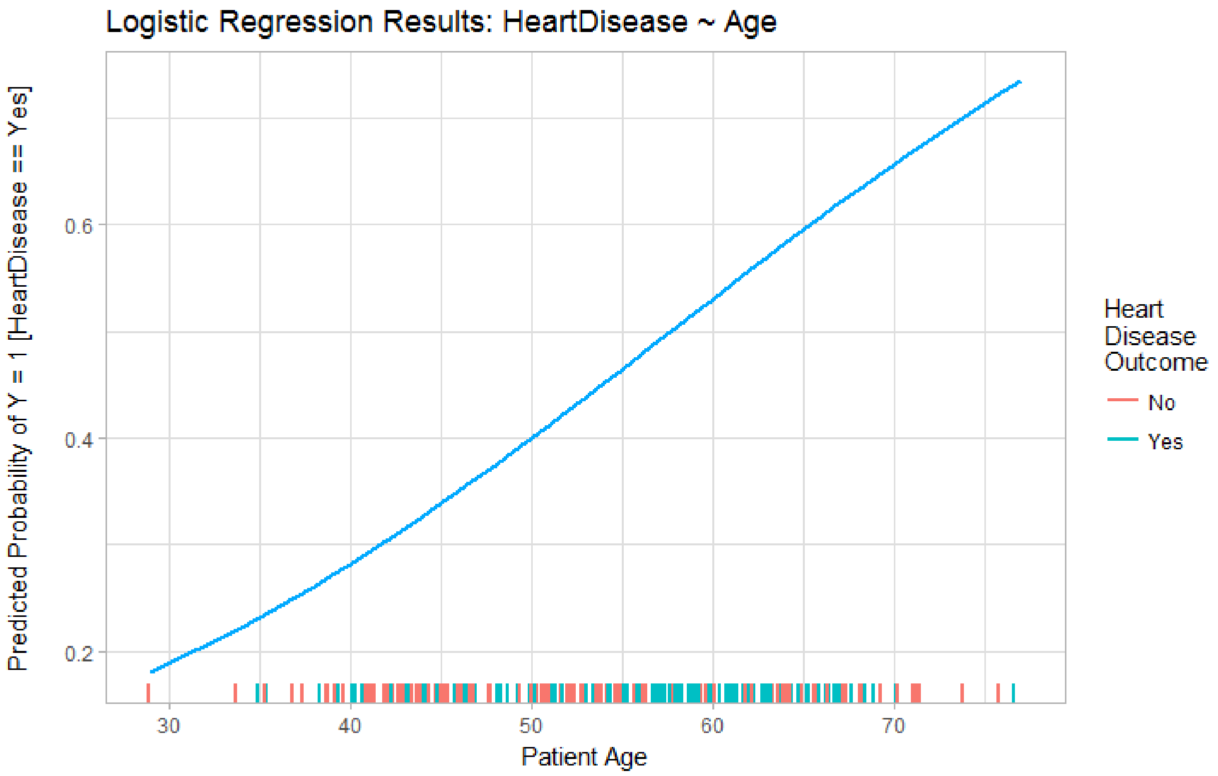
Logistic Regression
Review of the Model

In logistic regression, we directly modeled $P(Positive\ class \mid X = x)$ using the logistic function

The blue function on the right, is the estimated logistic function for the probability that a patient will develop heart disease as a function of the patient's age

Once we obtain estimated probabilities in logistic regression, we classify the categorical outcome variable based on a cut-off value (usually 0.5)

- This happens for all patients approximately 59 years and older based on the logistic curve on the right

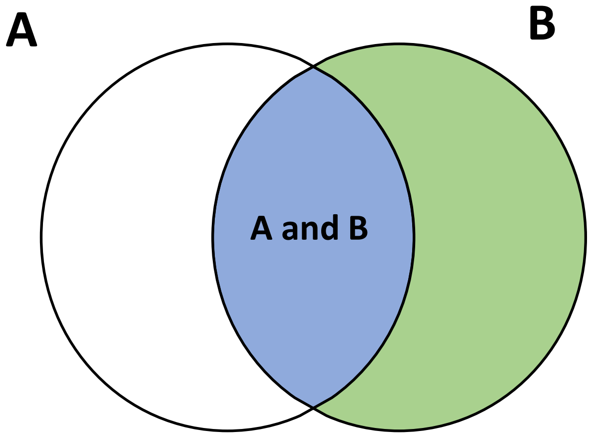


George Mason University School of Business

Probability
Conditional Probability

Conditional probability is the probability that event **A** occurs given that event **B** has already occurred

- In supervised learning, *classification* models are based on this concept



The notation for writing the probability of **A** *given* **B** is

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

Ratio of

- the probability that both **A** and **B** occur
- The probability that **B** occurs

George Mason University
School of Business

Probability

Conditional Probability

Suppose we are interested in knowing the probability of a fair die landing on 6 given that we know the outcome is an even number

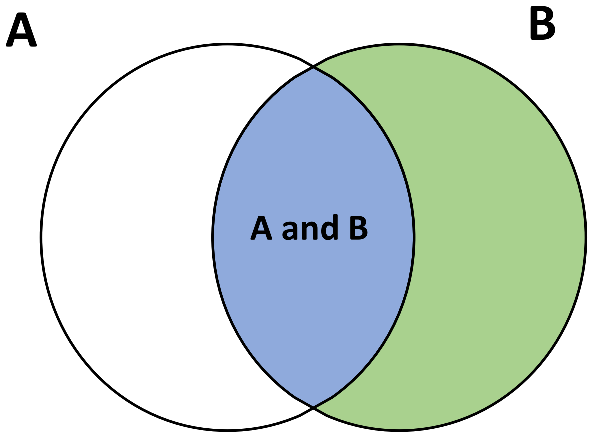
Sample Space

- Set of all possible outcomes
- 1, 2, 3, 4, 5, 6
- All equally likely with probability of 1/6

A = Land on 6

B = Land on an even number

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)} = \frac{1/6}{3/6} = \frac{1}{3}$$



GEORGE MASON UNIVERSITY
School of Business

Probability
Bayes Theorem

Bayes theorem describes the probability of an event based on prior knowledge and is useful for changing the order in a conditional probability statement

- Bayes' theorem is important in ***classification***
  - Allows us to estimate the probability that an observation is of a particular class given a predictor value ***based on the likelihood of the predictor value given that class***

$$P(A|B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)}$$

→

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

George Mason University School of Business

Probability

Bayes Theorem – An Example

Classification Task

Predict whether a patient will develop heart disease

A patient has a maximum heart rate of 140, what should we predict?

- Let **Y** be the event that a patient develops heart disease (heart_disease = "Yes")
- Let **N** be the event that a patient does not develop heart disease (heart_disease = "No")
- Let **B** be the event that a patient has a maximum heart rate of 140

Our Training Data

Assumed to be a random sample of patients

heart_disease	maximum_heart_rate
Yes	140
Yes	124
Yes	140
No	98
No	102
No	140
No	101

Let's use our training data to estimate the probability of developing heart disease

$$P(Y|B) = \frac{P(B|Y)\,P(Y)}{P(B)} = \frac{\left(\frac{2}{3}\right)\left(\frac{3}{7}\right)}{\frac{3}{7}} = \frac{2}{3}$$

$$P(N|B) = \frac{P(B|N)\,P(N)}{P(B)} = \frac{\left(\frac{1}{4}\right)\left(\frac{4}{7}\right)}{\frac{3}{7}} = \frac{1}{3}$$

GEORGE MASON UNIVERSITY
School of Business

Linear Discriminant Analysis
Using Bayes Theorem to Obtain $P(Y = y | X = x)$

Logistic regression

- $P(Y = y | X = x)$ is modeled directly with the logistic function

Discriminant analysis

- Models the distribution of the predictor variables in each class of the outcome variable
- Uses Bayes Theorem to flip things around in order to obtain $P(Y = y | X = x)$

Recall that we can use Bayes Theorem to write the following:

$$P(Y = y | X = x) = \frac{P(X = x | Y = y)\, P(Y = y)}{P(X = x)}$$

To make classification decisions, we can **ignore the denominator** since it is just a constant

- We ignored the 3/7 in the denominator in the previous example since we get the same outcome

All we need to estimate for each class of our response variable is

$$P(X = x | Y = y)\, P(Y = y)$$

heart_disease	maximum_heart_rate
Yes	140
Yes	124
Yes	140
No	98
No	102
No	140
No	101

GEORGE MASON UNIVERSITY
School of Business

Linear Discriminant Analysis

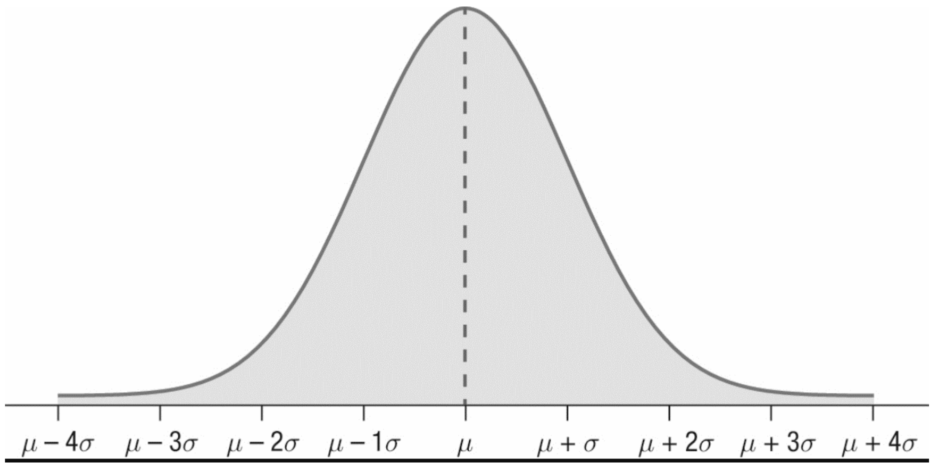
Bayes Theorem in Discriminant Analysis

$P(X = x | Y = y) \, P(Y = y)$ is written differently in the discriminant analysis setting:

$$\pi_k \;\; f_k(x)$$

$f_k(x)$

- Represents $P(X = x | Y = k)$
- Probability density function for the predictor variable X within class k
- $f_k(x)$ are assumed to be ***Normal distributions***



https://mat117.wisconsin.edu/wp-content/uploads/2014/12/Sec03.-NormalDis.png

π_k

- Represents $P(Y = k)$
- Prior probability that an observation is in class k

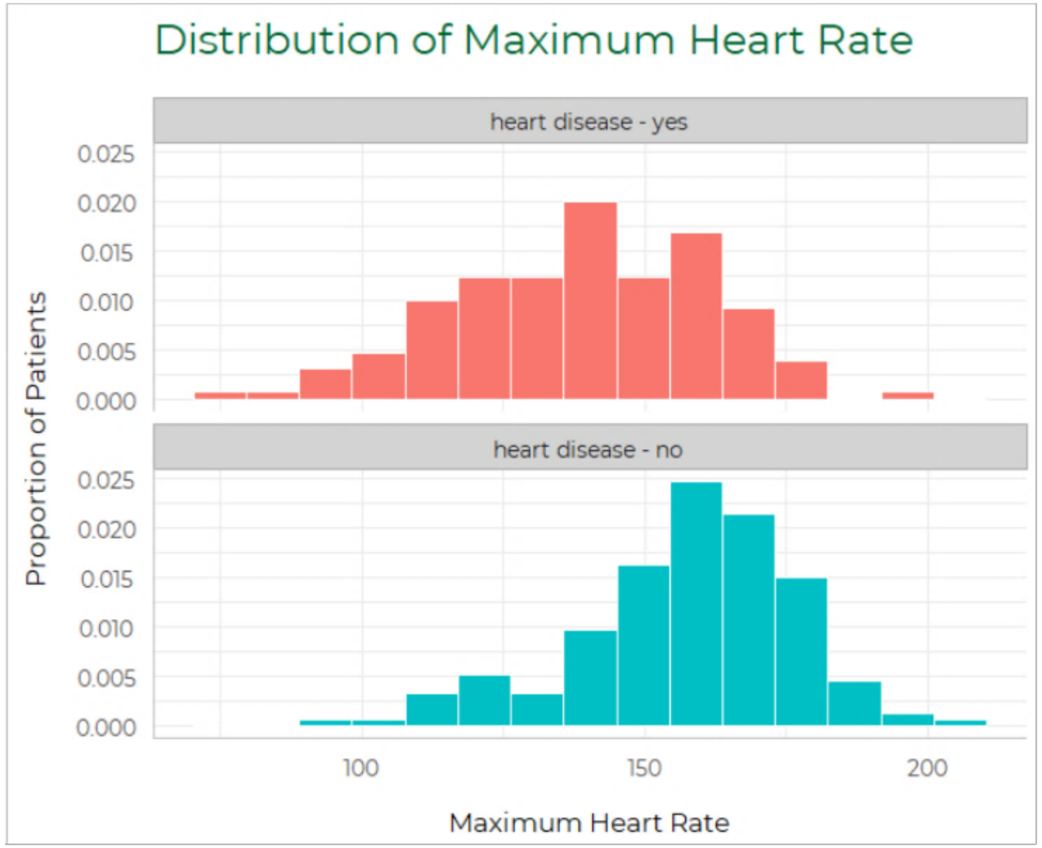
School of Business

Linear Discriminant Analysis
Goal of Discriminant Analysis

The goal of discriminant analysis is to estimate

$$\pi_k \ f_k(x)$$

for **each class** of the outcome variable and to predict the class with the largest estimated probability



GEORGE MASON UNIVERSITY
School of Business

Linear Discriminant Analysis

Linear Discriminant Analysis with One Predictor Variable

We assume that for each class of the outcome variable, the predictor variable follows a ***Normal distribution with common variance***

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}$$

To classify the outcome variable at the value $X = x$, we need to find the maximum value of $\pi_k \ f_k(x)$ among the different classes of the outcome variable

School of Business

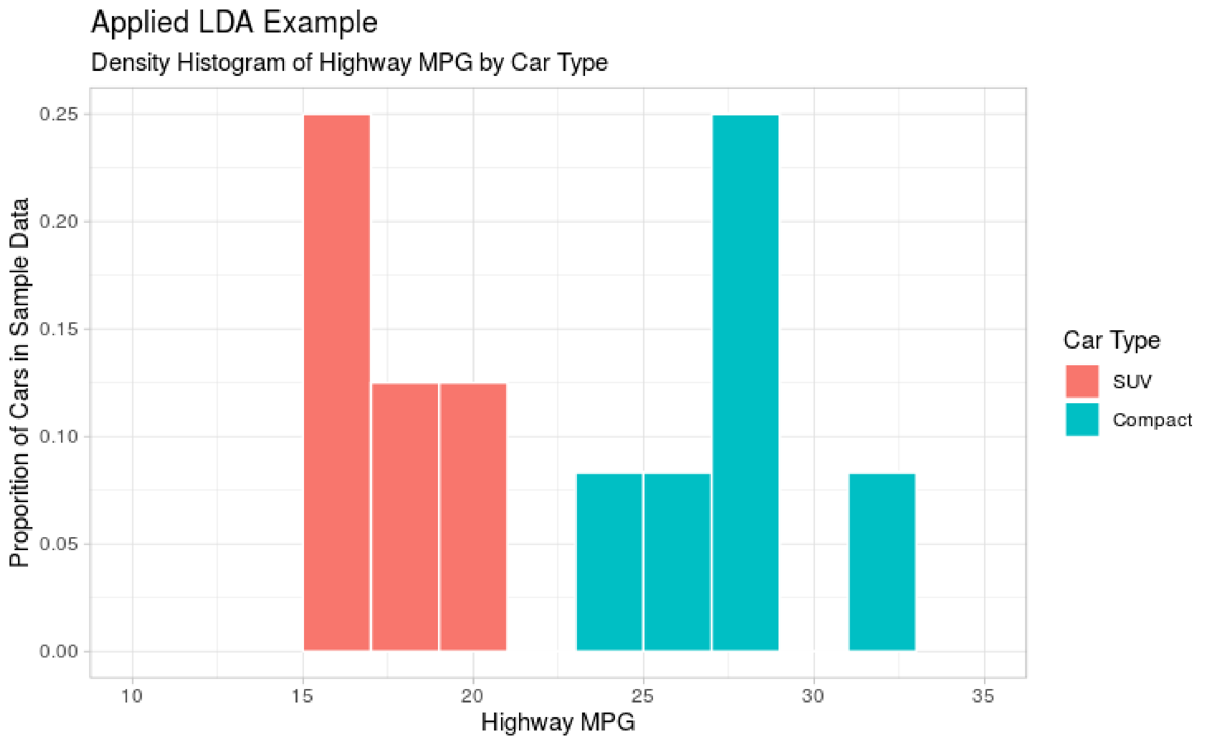
Linear Discriminant Analysis
Applied Example with One Predictor Variable

Goal - Use LDA to predict **Car Type** using **Highway MPG**

- For any new value of Highway MPG, we classify based on which of the quantities below is greater

$$\pi_{SUV} \ f_{SUV}(x) \quad \text{or} \quad \pi_{Compact} \ f_{Compact}(x)$$

Car Type	Highway MPG
SUV	20
SUV	16
SUV	19
SUV	16
Compact	32
Compact	27
Compact	29
Compact	28
Compact	28
Compact	25



Linear Discriminant Analysis

Estimating Parameters From the Data

Next, we estimate the mean and standard deviation of ***Highway MPG*** within each class of ***Car Type*** using the summary statistics $\bar{X}$ and S

Car Type	Highway MPG
SUV	20
SUV	16
SUV	19
SUV	16
Compact	32
Compact	27
Compact	29
Compact	28
Compact	28
Compact	25

→

Class k	Class Probability π_k	Class Average $\widehat{\mu_k}$	Class Variance $\widehat{\sigma^2}$	Class Standard Deviation $\hat{\sigma}$
SUV	4/10 = *0.4*	17.8	4.3	2.1
Compact	6/10 = *0.6*	28.2	5.4	2.3

GEORGE MASON UNIVERSITY
School of Business

Linear Discriminant Analysis

Estimating Common Variance From Groups – Pooled Variance

Class k	**Class Probability** π_k	**Class Average** $\widehat{\mu_k}$	**Class Variance** $\widehat{\sigma^2}$	**Class Standard Deviation** $\widehat{\sigma}$
SUV	4/10 = *0.4*	17.8	**4.3**	2.1
Compact	6/10 = *0.6*	28.2	**5.4**	2.3

Linear Discriminant Analysis assumes a ***common variance*** within all classes of the outcome variable

- We must combine our two estimates of the variance above into one overall estimate
- The *Pooled Sample Variance*, usually written as S_p^2, is used for this purpose
- S_p^2 is just a weighted average of the estimated group variances

Formula (n_i **is the number of observations in group *i*)**

$$S_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)\, s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{(4-1) * 4.3 + (6-1) * 5.4}{(4-1)+(6-1)} = \frac{(3) * 4.3 + (5) * 5.4}{8} = 4.988$$

Pooled Standard Deviation, $S_p = \sqrt{S_p^2} = \sqrt{4.988} = 2.2$

George Mason University School of Business

Linear Discriminant Analysis

Estimating Group-Specific Normal Distributions

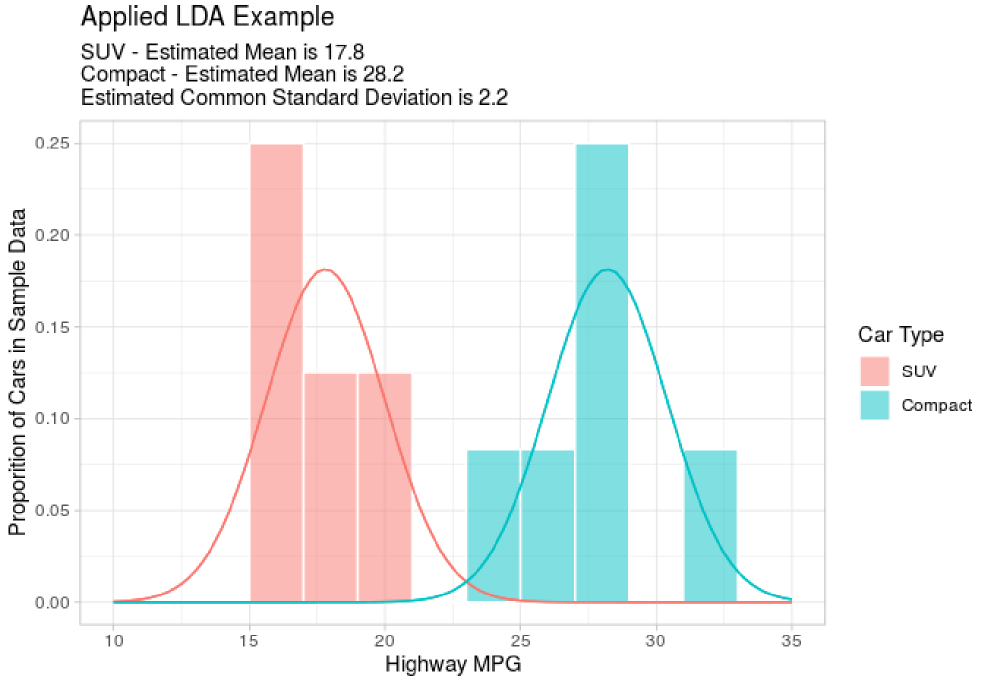
Class k	Class Probability π_k	Class Average $\widehat{\mu_k}$	Class Variance $\widehat{\sigma^2}$	Class Standard Deviation $\hat{\sigma}$	Pooled Standard Deviation, S_p
SUV	4/10 = *0.4*	**17.8**	4.3	2.1	**2.2**
Compact	6/10 = *0.6*	**28.2**	5.4	2.3	

Now we have all the estimates that we need

- Each group is modeled as a Normal distribution with group-specific mean and common standard deviation

- $f_{SUV}(x) = \frac{1}{\sqrt{2\pi}\,(2.2)} e^{\frac{-1}{2}\left(\frac{x-17.8}{2.2}\right)^2}$

- $f_{Compact}(x) = \frac{1}{\sqrt{2\pi}\,(2.2)} e^{\frac{-1}{2}\left(\frac{x-28.2}{2.2}\right)^2}$



Linear Discriminant Analysis
Predictions For New Data

New Data: Highway MPG is ***26***

What type of car is it?

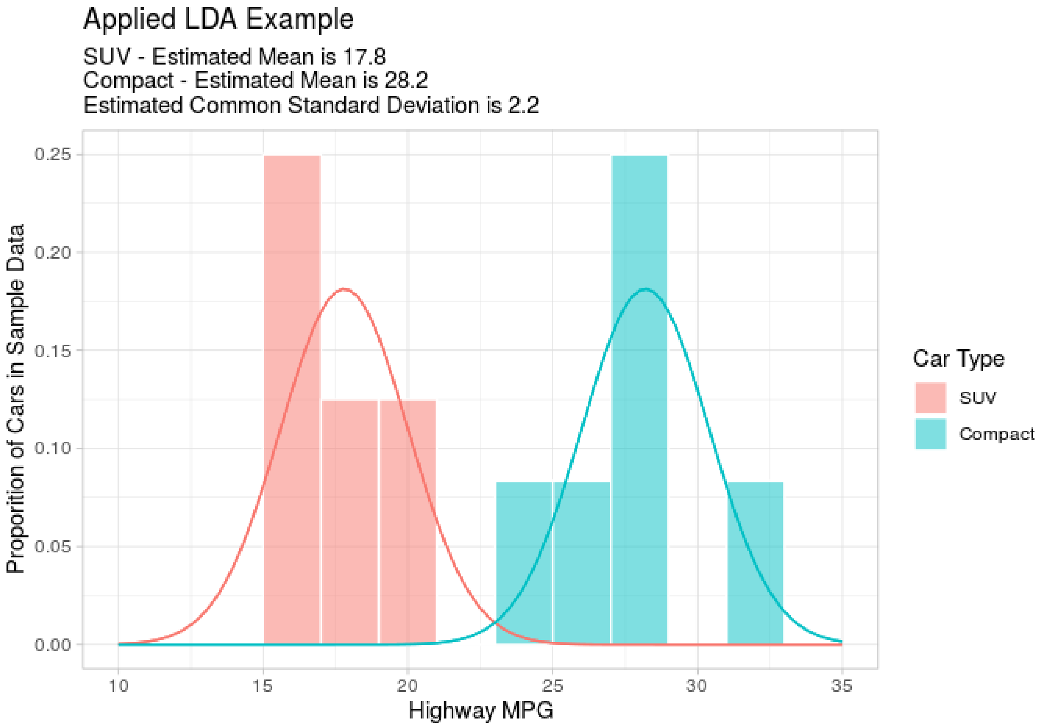
- $\pi_{SUV}\ f_{SUV}(\mathbf{26}) =$

$$(0.4) * \frac{1}{\sqrt{2\pi}\,(\mathbf{2.2})} e^{\frac{-1}{2}\left(\frac{\mathbf{26}-\mathbf{17.8}}{\mathbf{2.2}}\right)^2} = 0.00015$$

- $\pi_{Compact}\ f_{Compact}(\mathbf{26}) =$

$$(0.6) * \frac{1}{\sqrt{2\pi}\,(\mathbf{2.2})} e^{\frac{-1}{2}\left(\frac{\mathbf{26}-\mathbf{28.2}}{\mathbf{2.2}}\right)^2} = \mathbf{0.145}$$

- Conclusion – we predict **Compact**



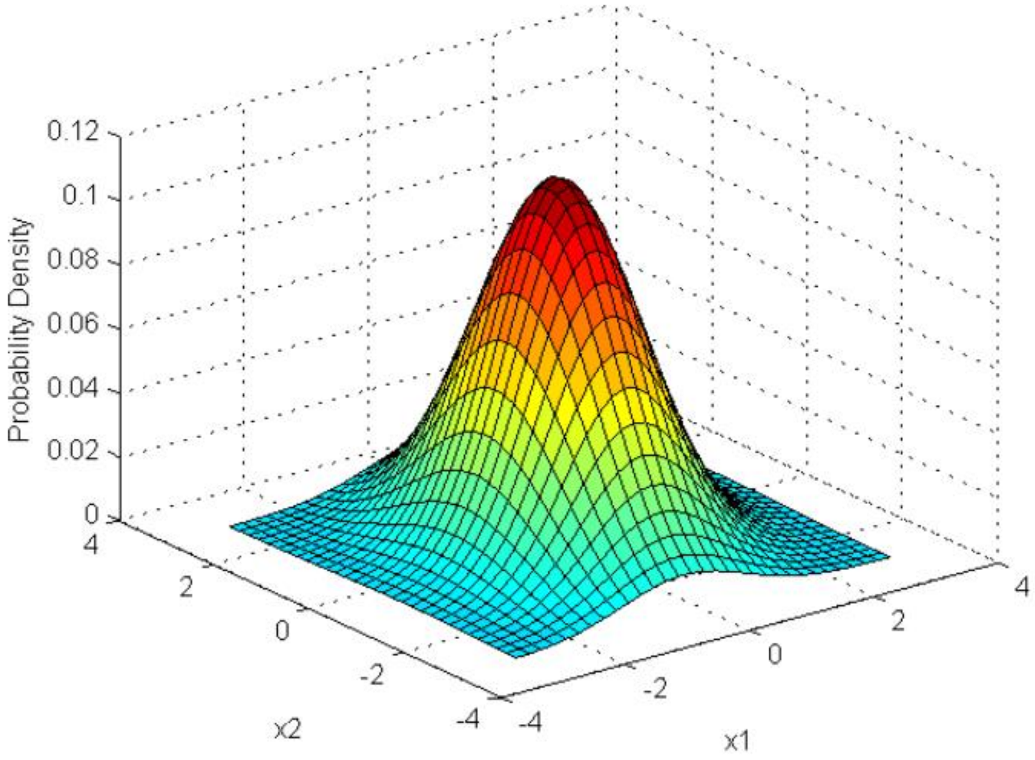
School of Business

Linear Discriminant Analysis

Linear Discriminant Analysis with Multiple Predictor Variables

Linear discriminant analysis is easily extended to the case in which we have multiple numeric predictor variables

- In this case, the set of predictor variables are assumed to follow a ***multivariate normal distribution*** with common covariance matrix within each class



GEORGE MASON UNIVERSITY

School of Business

Quadratic Discriminant Analysis

Quadratic Discriminant Analysis with One Predictor Variable

In Quadratic Discriminant Analysis (QDA) we assume that for each class of the outcome variable, the predictor variable follows a ***Normal distribution with class specific variance***

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} e^{\frac{-1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2}$$

As in LDA, we need to find the maximum value of $\pi_k \, f_k(x)$ among the different classes of the outcome variable

GEORGE MASON UNIVERSITY
School of Business

Discriminant Analysis
LDA – Linear Decision Boundary

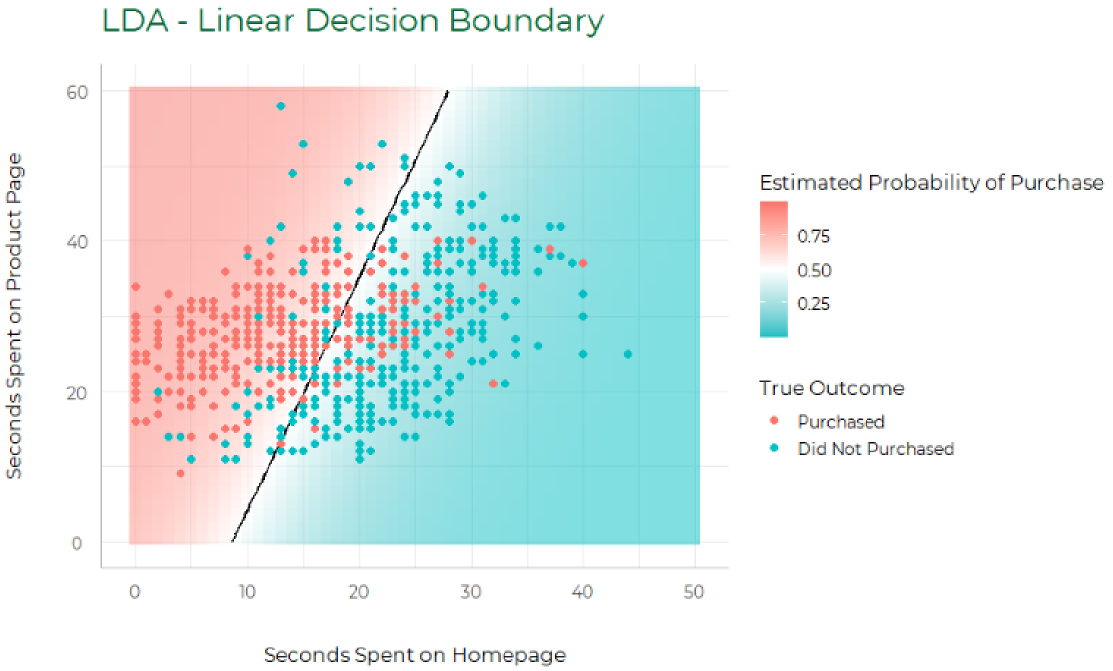
Example

Predict whether a customer will purchase a product based on the seconds they have spent browsing a company's homepage and product page

Segmenting the predictor values into distinct, non-overlapping regions to predict a category

Outcome — Predictors

Outcome	Seconds Homepage	Seconds Product Page
Did Not Purchase	4	30
Purchased	32	43
Did Not Purchase	2	22
Purchased	24	36



School of Business

Discriminant Analysis
QDA – Quadratic Decision Boundary

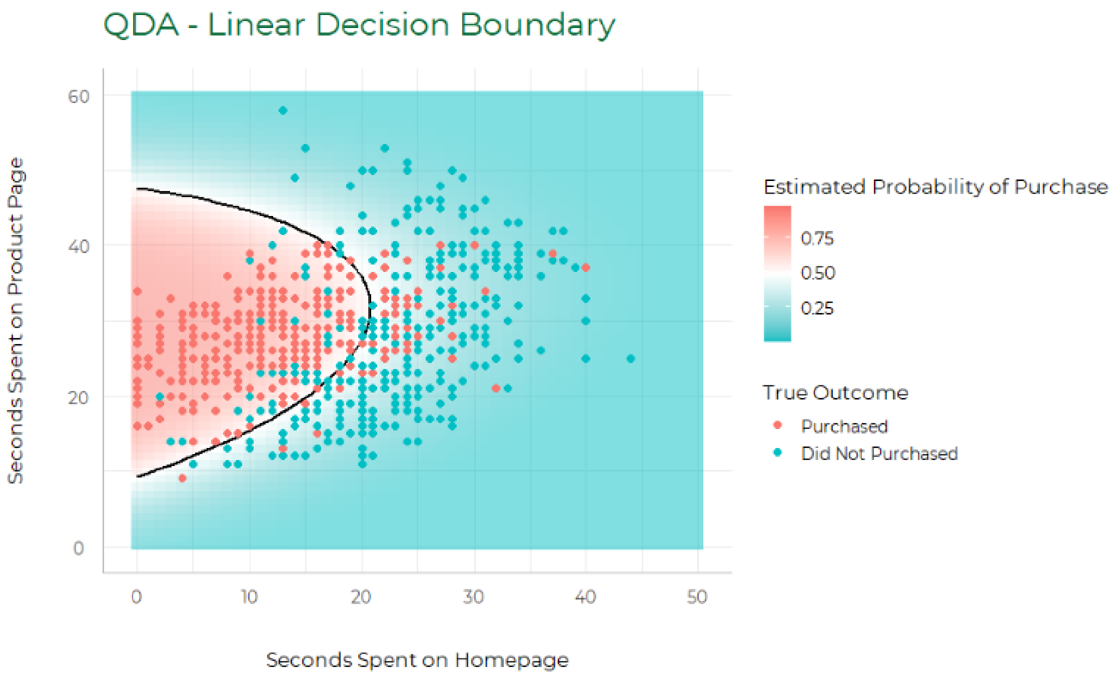
Example

Predict whether a customer will purchase a product based on the seconds they have spent browsing a company's homepage and product page

Segmenting the predictor values into distinct, non-overlapping regions to predict a category

Outcome | Predictors

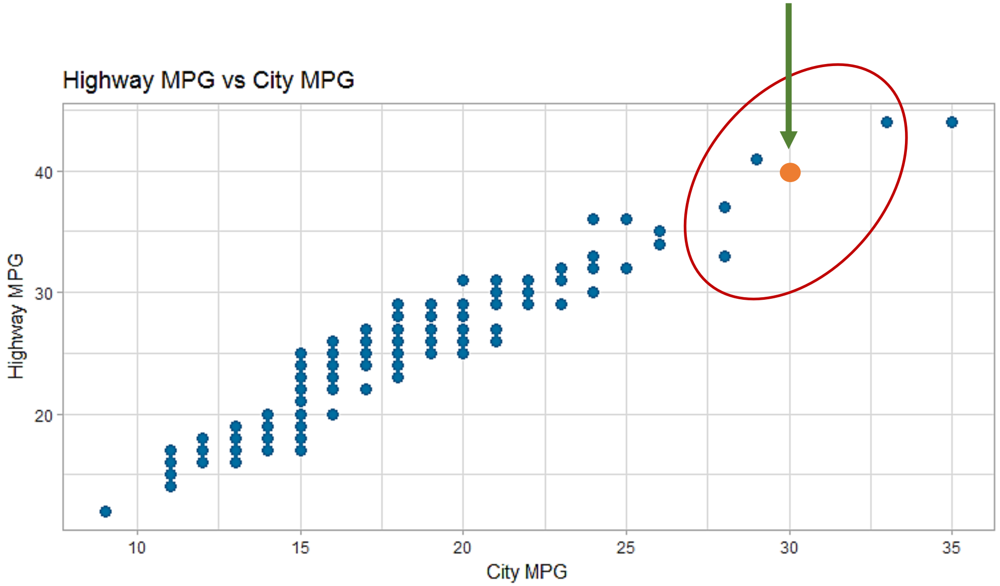
Outcome	Seconds Homepage	Seconds Product Page
Did Not Purchase	4	30
Purchased	32	43
Did Not Purchase	2	22
Purchased	24	36



K-Nearest Neighbor (*KNN*)

A Non-Parametric Approach to Estimating an Outcome Variable

- *KNN* is simple non-parametric technique that can be applied to both regression and classification problems
- *KNN* uses a simple approach in making predictions
  - Finds the *K* ***nearest*** points to a particular predictor variable value and predicts either the mean of these points (regression) or the outcome class with the largest proportion in the sample of *K* points
- To find the *K* nearest points, most *KNN* algorithms use **Euclidean distance** by default



cty	hwy	fl	class
26	35	r	compact
28	33	[illegible]	subcompact
28	37	[illegible]	compact
29	41	[illegible]	subcompact
33	44	[illegible]	compact
35	44	d	subcompact

Predict highway MPG for city MPG of 30 with *K = 4*

$$\frac{33 + 37 + 41 + 44}{4} = 38.8$$

School of Business

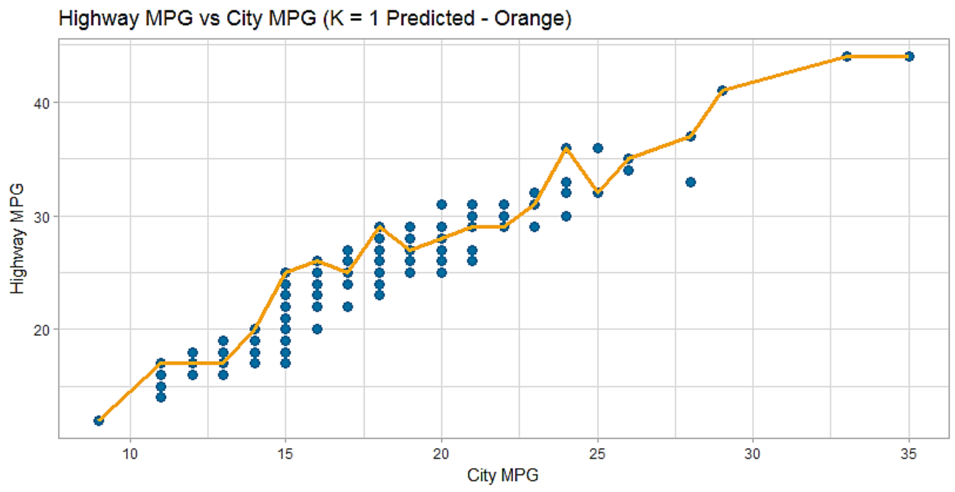
K-Nearest Neighbor
Model hyperparameters

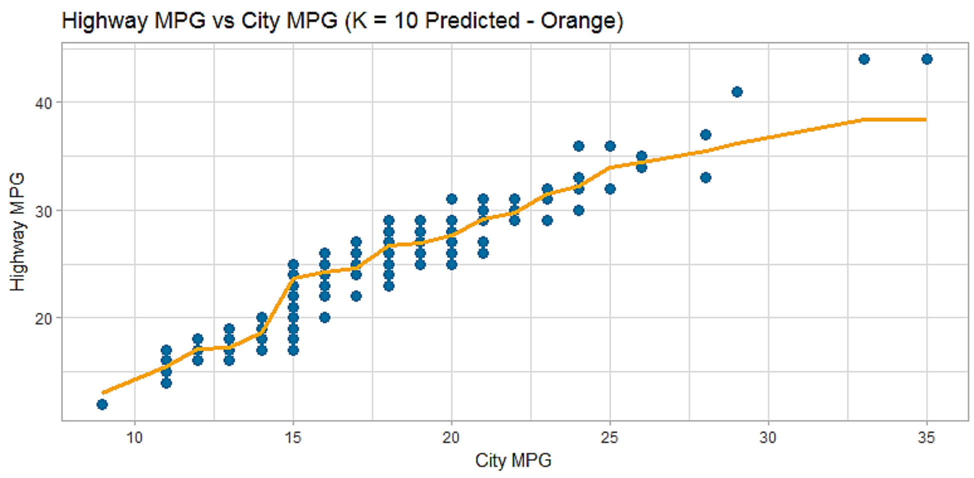
K is known as a model **hyperparameter**

As we increase the value of *K*, the resulting predictions become more smooth

The challenge is to find the optimal value of *K* that produces the lowest prediction error

- Known as *hyperparameter tuning*
- Accomplished with the **tune** package from **tidymodels**





George Mason University School of Business

Machine Learning Process

Response	Predictor 1	Predictor 2
Yes	12.4	Low
No	14.2	Moderate
Yes	16.8	High
…	…	…

70%

30%

Training Set

Response	Predictor 1	Predictor 2
Yes	12.4	Low
Yes	16.8	High
No	11.8	Low
…	…	…

Test Set

Response	Predictor 1	Predictor 2
Yes	11.2	Low
No	14.2	Moderate
…	…	…

Model Training

Trained Model

Feature Engineering

Apply transformations
(learned on training data)

Predictions
Accuracy Assessment

School of Business

Hyperparameter Tuning

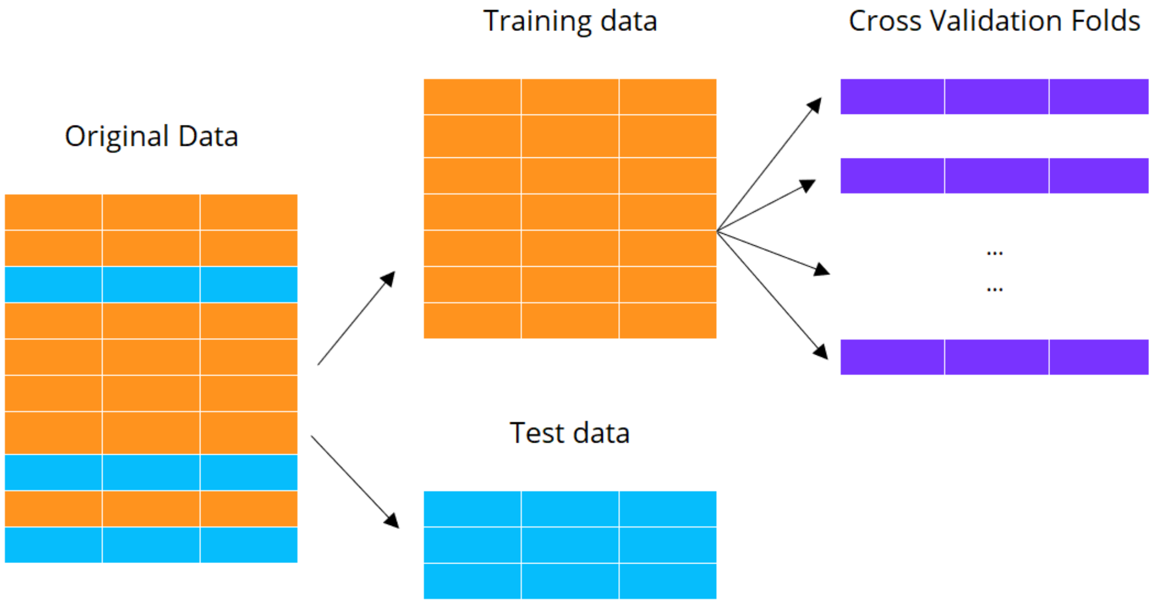
K-fold Cross Validation

There are drawbacks to the training/test set approach

- We only get one estimate of model performance (on the test set)

***K*-fold cross validation** is one way to improve our estimates

- Randomly divide the training data into *K* equal-sized parts

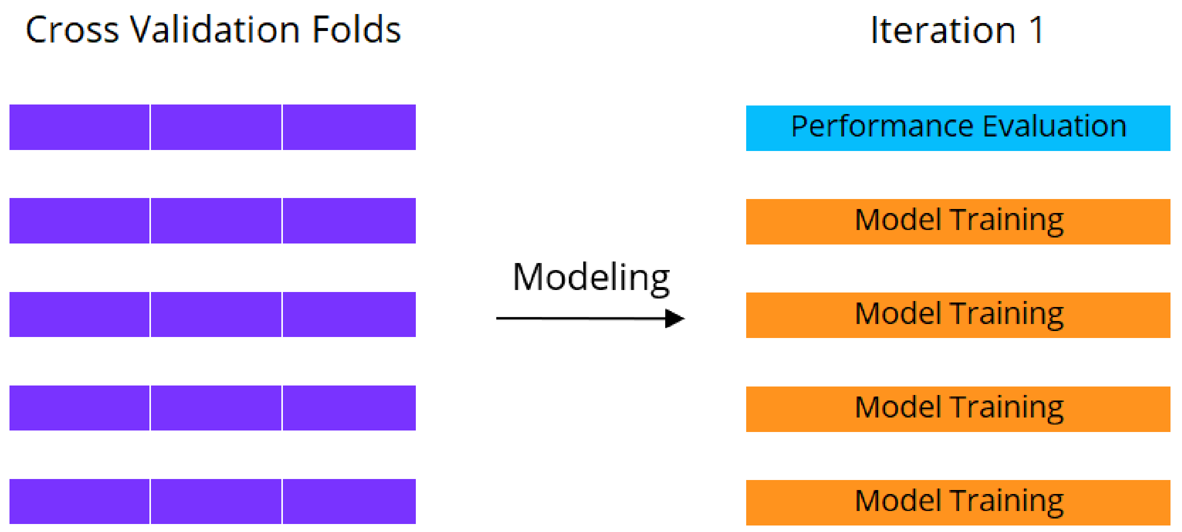


Hyperparameter Tuning

K-fold Cross Validation

For each *K*

- Leave out data set *K* and fit the model to the other combined *K* − 1 data sets

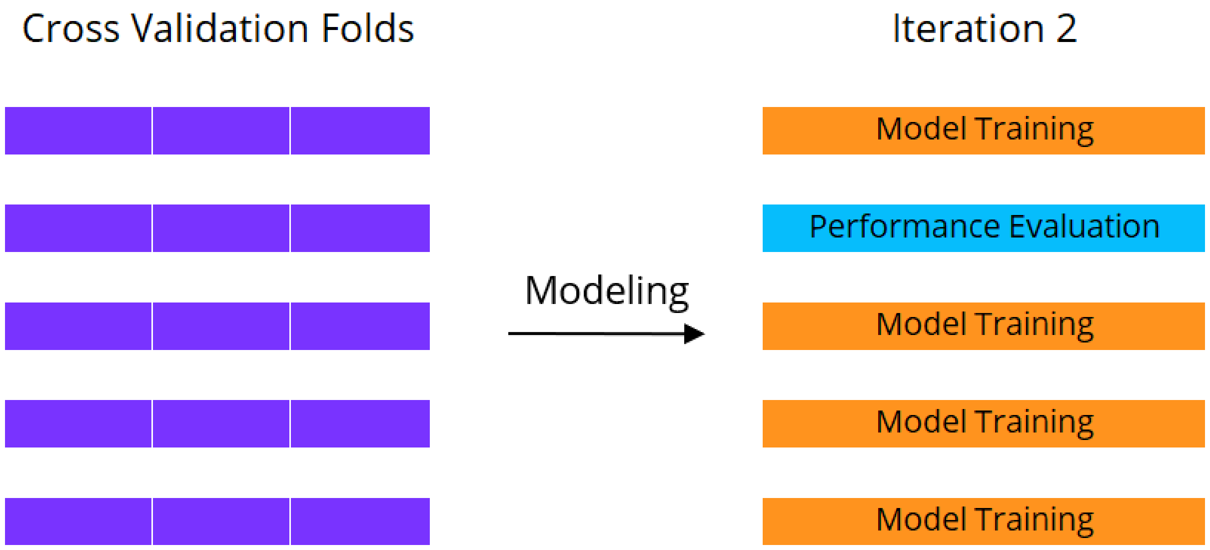


Hyperparameter Tuning

K-fold Cross Validation

For each *K*

- Leave out data set *K* and fit the model to the other combined $K - 1$ data sets
- Repeat this process for various values of our hyperparameters



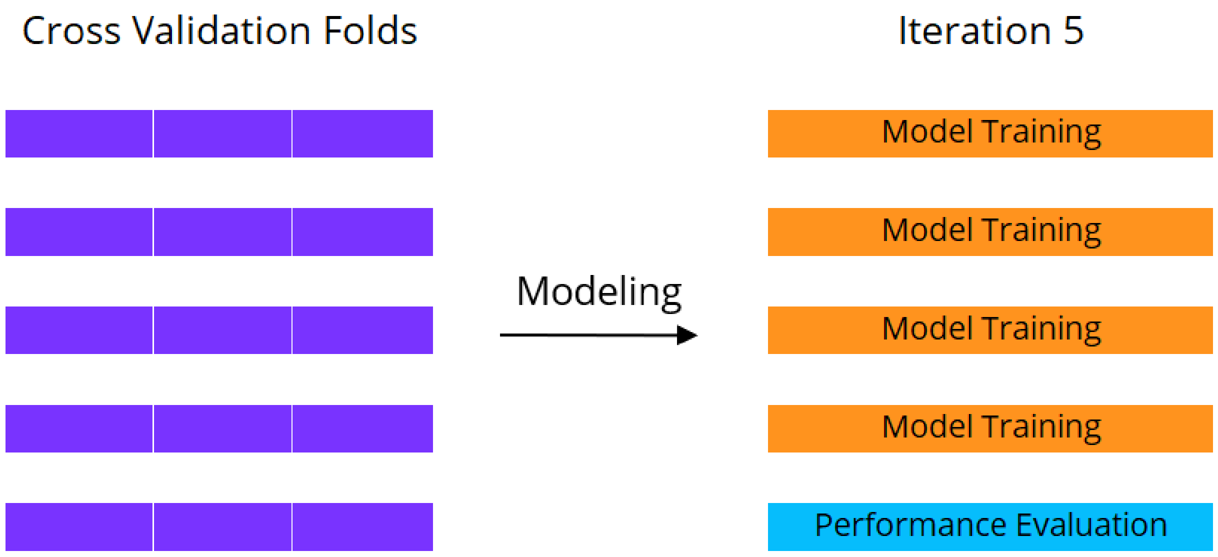
Hyperparameter Tuning

K-fold Cross Validation

For each *K*

- Leave out data set *K* and fit the model to the other combined $K - 1$ data sets
- Repeat this process for various values of our hyperparameters
- Select the best hyperparameter value(s) based on cross validation results

Neighbors (K)	Fold	ROC AUC
5	1	0.74
5	2	0.68
…	…	…
10	1	0.59
…	…	…
25	5	0.87



School of Business