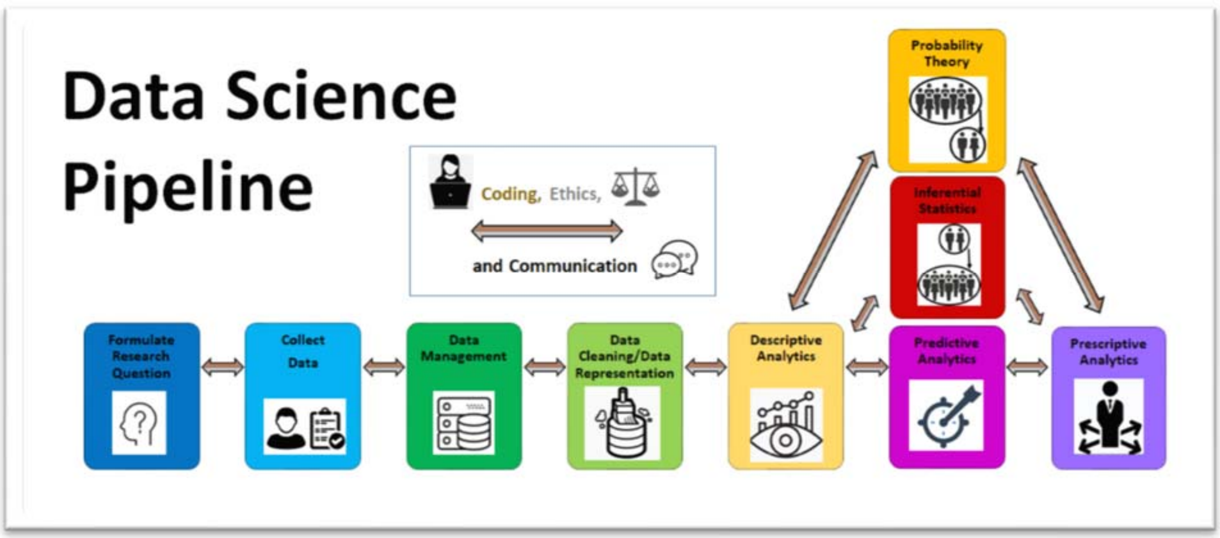
Unit 18: Training Data vs. Test Data

Case Studies:

- To introduce the concept of **using training data to *build* a model** and **using test data to *test* a model for it's predictive capabilities** we will, again, examine the relationship between a:
  - **Categorical response variable:** support for a certain opinion (favor/not in favor) and an
  - **Explanatory variables:**
    - Sex
    - Party, and
    - Age



Summary of Concepts:

1. **Different Goals for Building a Regression Model**
2. **Problem with Overfitting a Regression Model**
3. **Training vs. Test Dataset**
4. **Case Study:** Building a Model that is Good at Predicting Approval for the President's Foreign Policy with Age, Sex, and Political Affiliation *with New Data*

1. DIFFERENT GOALS FOR BUILDING A REGRESSION MODEL

Data

Suppose you work at a data science firm and we have access to the **Body Dimensions dataset** that we have used in the past that is comprised of various body measurements of a random sample of healthy adults.

	bicep_girth	age	sex	weight	height
0	32.5	21	Male	65.6	174.0
1	34.4	23	Male	71.8	175.3
2	33.4	28	Male	80.7	193.5
3	31.0	23	Male	72.6	186.5
4	32.0	22	Male	78.8	187.2
...	...	...	...	...	...
482	30.3	29	Female	71.8	176.5
483	30.1	21	Female	55.5	164.4
484	27.4	33	Female	48.6	160.7
485	30.6	33	Female	66.4	174.0
486	33.2	38	Female	67.3	163.8

You have two clients who would like your help to meet the following goals.

Clients

Client 1 Goal: This client works in a U.S. public health agency and is interested in **understanding the relationship** between bicep girth, age, sex, weight, and height of ALL healthy adults. Having in this information can lead to better informed policies surrounding muscle mass development.

Client 2 Goal: This client works at a clothing company whose goal is to design and ship well-fitted business jackets to customers given their age, sex, weight, and height that they fill out in a survey. One important aspect of producing a well-fitted business jacket is knowing the bicep girth of the customer, however most customers do not know their bicep girth. Therefore, being able to **accurately predict** the bicep girth of a customer given the information that they supply is very important to this client.

Strategies for Building a Model with this Data

Which of the following model building strategies would you suggest for each client?

Strategy 1: Give the client the linear regression model that only **contains the slopes that are statistically significant**.

ie. $\widehat{bicep\ girth} = 30.7279 + 3.3844 sex[T.Male] + 0.2449 weight - 0.1060 height$

OLS Regression Results

Dep. Variable:	bicep_girth	**R-squared:**	0.831
Model:	OLS	**Adj. R-squared:**	0.829
Method:	Least Squares	**F-statistic:**	590.9
Date:	Wed, 21 Apr 2021	**Prob (F-statistic):**	2.94e-184
Time:	20:38:31	**Log-Likelihood:**	-963.88
No. Observations:	487	**AIC:**	1938.
Df Residuals:	482	**BIC:**	1959.
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	31.4253	2.032	15.465	0.000	27.432	35.418
sex[T.Male]	3.4235	0.235	14.590	0.000	2.962	3.885
age	-0.0132	0.009	-1.547	0.123	-0.030	0.004
weight	0.2475	0.009	26.789	0.000	0.229	0.266
height	-0.1088	0.013	-8.129	0.000	-0.135	-0.083

Omnibus:	13.978	**Durbin-Watson:**	1.993
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	15.394
Skew:	0.347	**Prob(JB):**	0.000454
Kurtosis:	3.526	**Cond. No.**	4.78e+03

OLS Regression Results

Dep. Variable:	bicep_girth	**R-squared:**	0.830
Model:	OLS	**Adj. R-squared:**	0.829
Method:	Least Squares	**F-statistic:**	784.7
Date:	Wed, 21 Apr 2021	**Prob (F-statistic):**	3.19e-185
Time:	20:58:11	**Log-Likelihood:**	-965.09
No. Observations:	487	**AIC:**	1938.
Df Residuals:	483	**BIC:**	1955.
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	30.7279	1.984	15.486	0.000	26.829	34.627
sex[T.Male]	3.3844	0.234	14.487	0.000	2.925	3.843
weight	0.2449	0.009	26.922	0.000	0.227	0.263
height	-0.1060	0.013	-7.980	0.000	-0.132	-0.080

Omnibus:	14.566	**Durbin-Watson:**	1.991
Prob(Omnibus):	0.001	**Jarque-Bera (JB):**	16.497
Skew:	0.345	**Prob(JB):**	0.000262
Kurtosis:	3.581	**Cond. No.**	4.60e+03

Strategy 2: Choose the best combination of explanatory variables (from sex, age, weight, and height) that will give the **best bicep girth predictions** for new customers (ie. not the people already in this dataset of 487 health adults).

2. PROBLEM WITH OVERFITTING A REGRESSION MODEL

Suppose we build two classifier models using the **given dataset** below. We call the dataset the **training data.**

Suppose this dataset is comprised of a random sample of **50 *actual positives*** (ie. observations with a response variable of 1) from a population of positives and a random sample of **50 *actual negatives*** (ie. observations with a response variable value of 0) from a population of negatives.

For each classifier model we also select a **prediction threshold** (shown in red below) with a rule that determines when/how we classify a given point as a 1 or a 0.

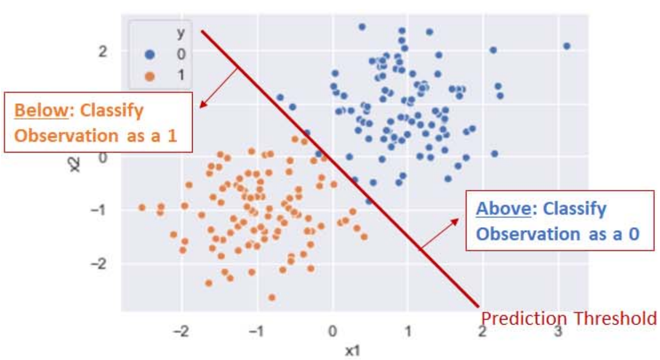
Training Data

a. What is the false positive rate and true positive rate of classifier model 1 (using the given prediction threshold)?

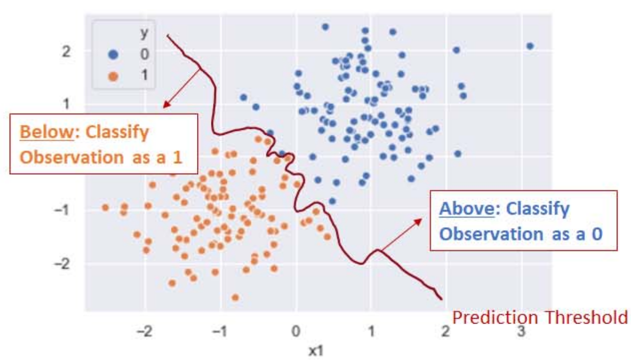
b. What is the false positive rate and true positive rate of classifier model 2 (using the given prediction threshold)?

c. If our goal is to classify the observations in **this training dataset** as accurately as possible, which model and threshold is better?

Classifier Model 1



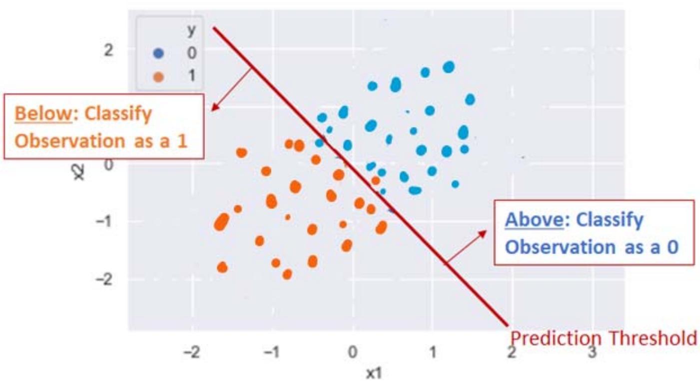
Classifier Model 2



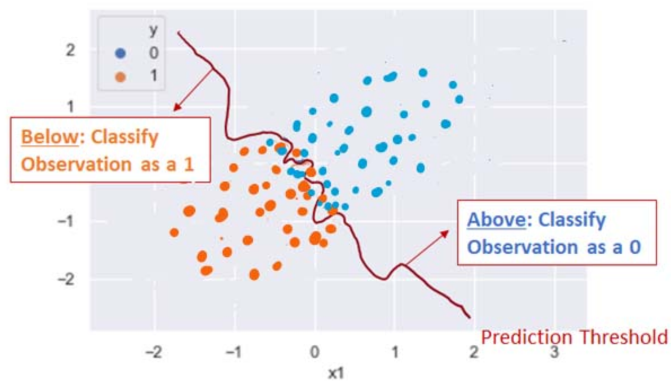
Test Data

d. Now suppose that we select *another* random sample of positives from the population of positives and *another* random sample of negatives from the population of negatives. We will call this new dataset the **test dataset**. We then classify these new points with the same two prediction thresholds shown above. If our goal is to classify the observations in **this test dataset** as accurately as possible, which model and threshold is better?

Classifier Model 1



Classifier Model 2



Definition of Overfitting

The example above introduces the concept of ______________________ a model to a given ______________________. This is a situation that arises in which we make decisions when fitting a model (and picking threshold) with a given ____________________________ that give us really good prediction accuracy for the ______________________. However, the model and threshold fit the ___________________ so well that when we try to make predictions with a new dataset the prediction accuracy is __________________.

Common Way to Overfit a Model

A common way to overfit your model is to create a model that has too many ____________________________.

3. TRAINING VS. TEST DATA

Definition of Training Dataset

Specifically, when fitting a linear regression model or a logistic regression model, we call the

_________________________ the dataset that was used to find the optimal values of $\widehat{\beta_0}, \widehat{\beta_1}, \dots, \widehat{\beta_p}$ in the

model. For the resulting model, we say that this model has been _____________ with the training dataset.

Problem

In order to train a linear or logistic regression model, we need to know the ___________________ values y

to determine how well our predictions were.

However, if our goal is to have the best __________________ predictions for **new datasets**, we often do

not know what the response variable values are.

So how are we supposed to get an idea of how well our trained linear or logistic regression models will do

with new data that doesn't have ____________________________?

Solution:

In order to solve this problem, we can take a dataset that we have where we know

_________________________, and **randomly** split it into two datasets:

1. The **training dataset**

   This dataset is used to ____________________________________.

2. The **test dataset**

   This dataset is used to ____________________________________.

4. <u>Case Study</u>: Building a Model that is Good at Predicting Approval for the President's Foreign Policy with Age, Sex, and Political Affiliation <u>with New Data</u>

<u>Goal:</u> Suppose we work at a political advertising agency. Rather than seek to **understand the relationship** between approval for the president's foreign policy with sex, age, and political affiliation, we would like build a model that will give us the **best predictions** for adults living in the U.S. in which *we don't know what they think about the president's foreign policy.*

<u>Data:</u>We can assume that this agency has the age, sex, political affiliation, and address of all registered voters in the state.

<u>Actions:</u> So one goal that this political advertising agency might have is to use this data to make predictions about whether a given person that lives at a particular house approves of the president's foreign policy. They could then use that information to decide whether to mail political advertising pamphplets to this address.

Go to the Unit 18 notebook section 4 to explore this case study.