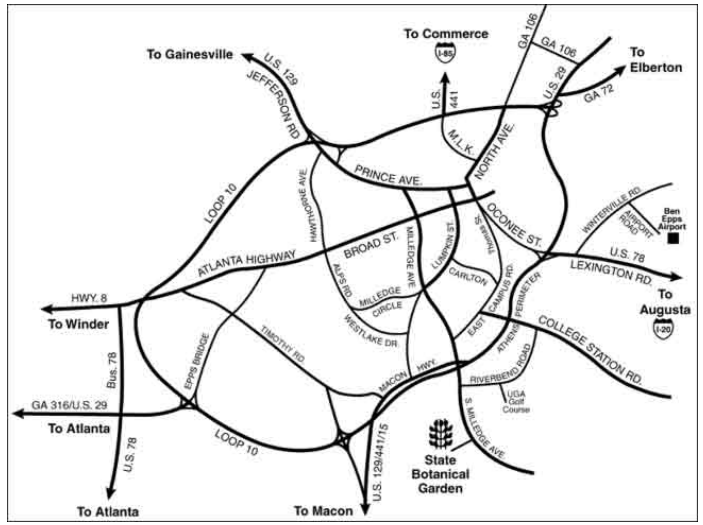
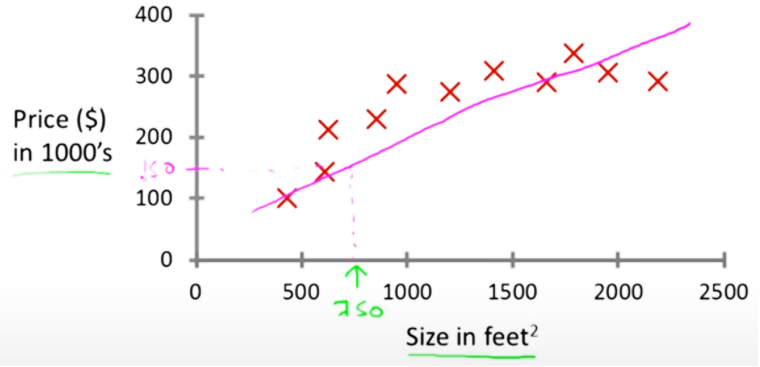
Model Basics

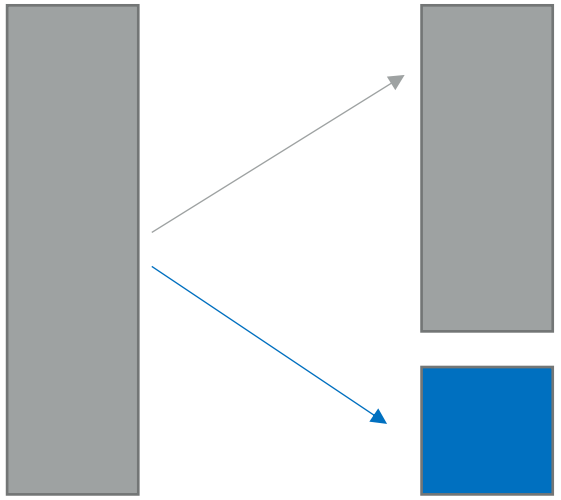
Carolina A. de Lima Salge
Assistant Professor
Terry College of Business
University of Georgia

Business Intelligence
Spring 2021

Terry College of Business
UNIVERSITY OF GEORGIA







Data

Training Data

Testing Data

Attributes

Target attribute

Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>**
Class label (value of Target attribute) is **no**

Model Defined

A **simplified* representation** of reality created for a **specific purpose**

- *based on some assumptions about what is and is not important, or sometimes based on constraints on information or tractability

Model Goal

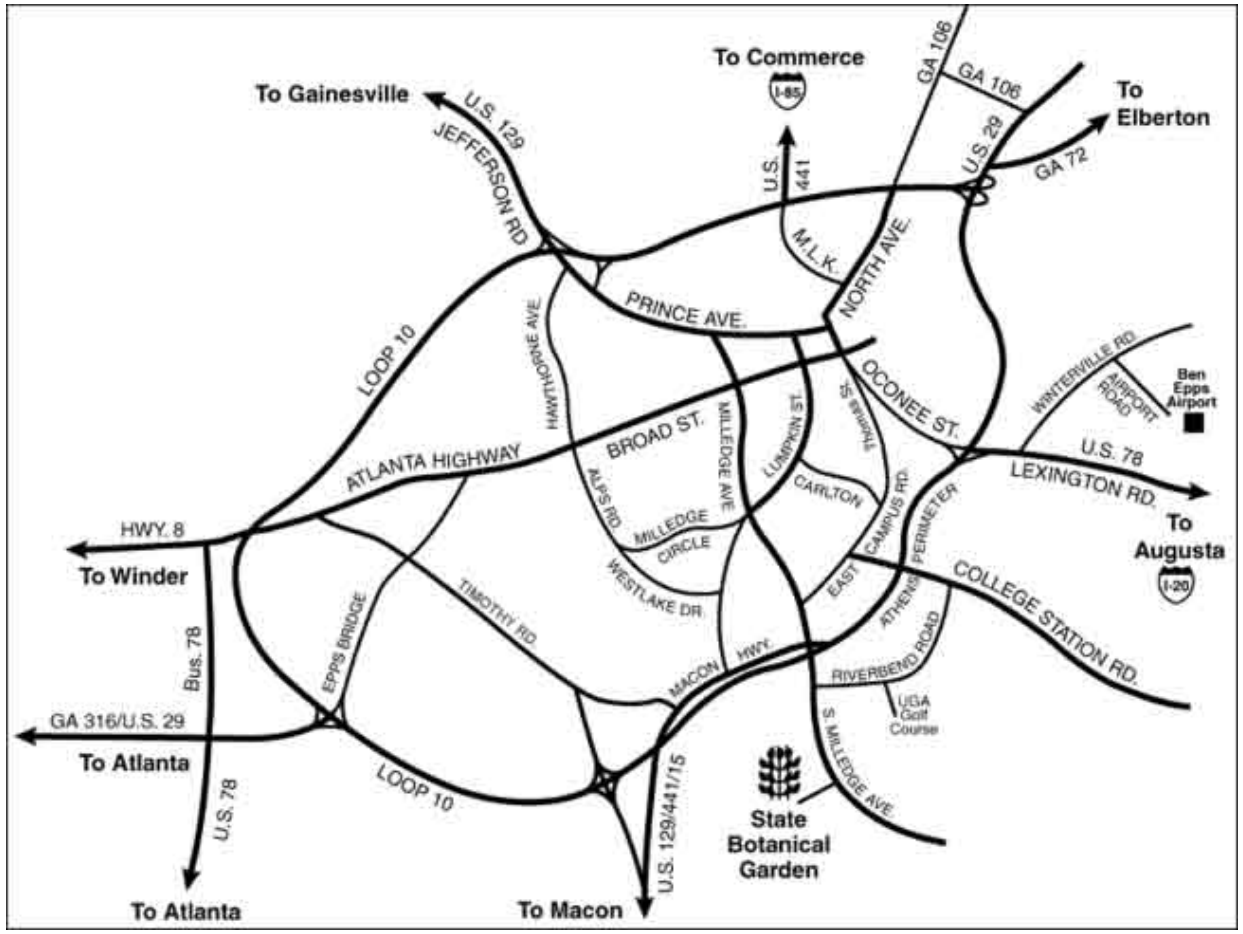
Not to uncover truth, but to discover a simple approximation that is still useful

- i.e., capture **true "signals"** (or patterns generated by the phenomenon of interest) and **ignore "noise"** (or random variation that you're not interested in)

“All models are wrong, but some are useful”

George Box

Model Example



Predictive (or Supervised Learning) Model

A formula for estimating the unknown value of interest:

- **The target!**
  - The formula could be mathematical or a logical statement, such as a rule. Often, it is a hybrid

More Terminology

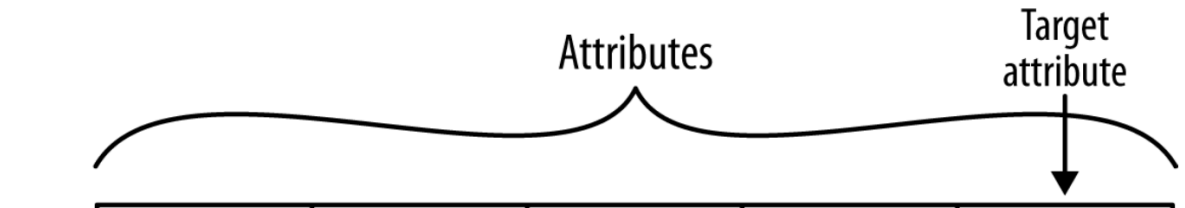
Instance / example = a fact or data point described by a set of attributes (also known as variables, columns, or features)

Model induction = the creation of models from data

Training data = the input data used for model induction

Testing data = the input data used for model testing

More Terminology



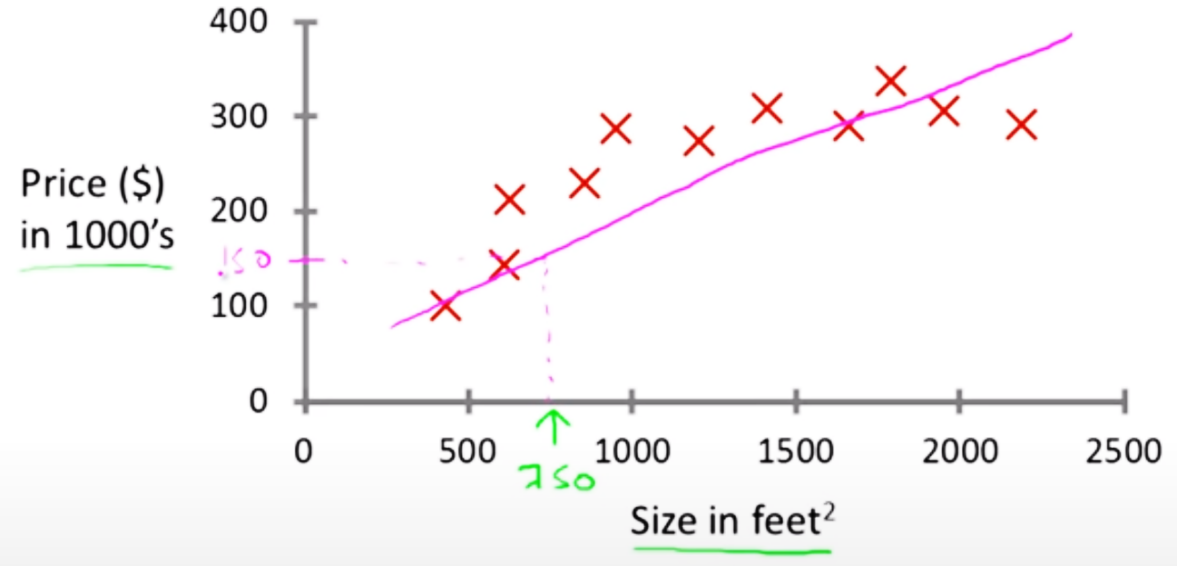
Name	Balance	Age	Employed	Write-off
Mike	$200,000	42	no	yes
Mary	$35,000	33	yes	no
Claudio	$115,000	40	no	no
Robert	$29,000	23	yes	yes
Dora	$72,000	31	no	no

This is one row (example).
Feature vector is: **<Claudio,115000,40,no>** <class, number, number, class>
Class label (value of Target attribute) is **no**

Predictive Model: Regression

Suppose you want to predict house prices, and you have some data about the price of a house (in thousands of $) over size (sqft)

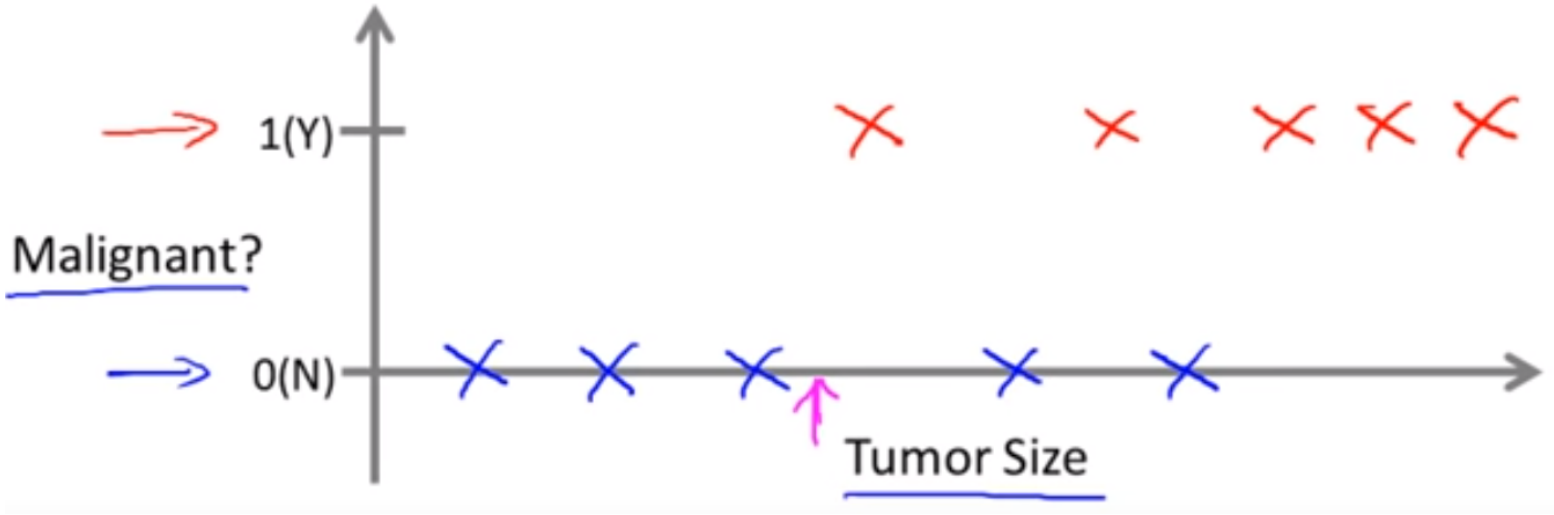
Estimate numeric value (e.g., with a linear regression)



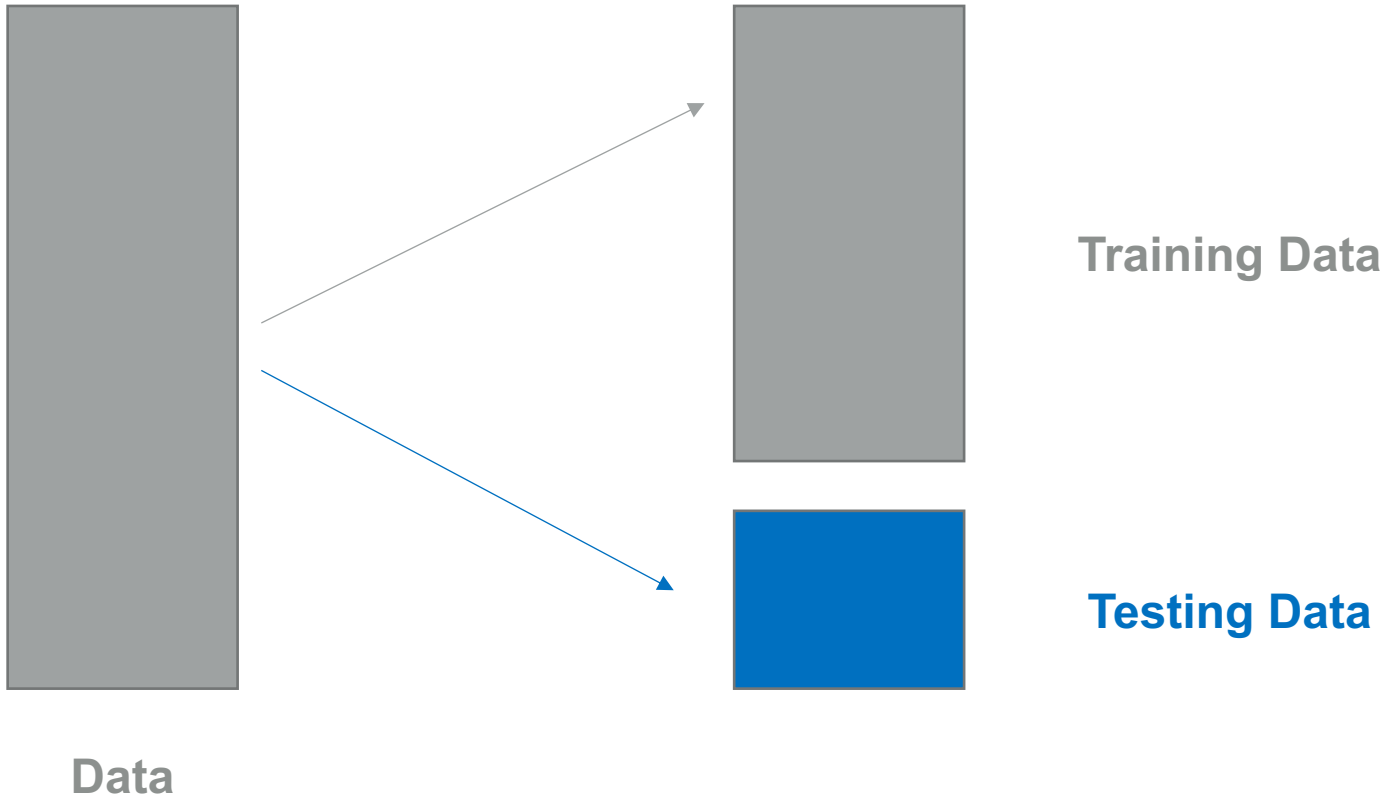
Predictive Model: Classification

Suppose you want to predict whether someone's breast cancer is malignant

Estimate class probability (e.g., with a logistic regression)



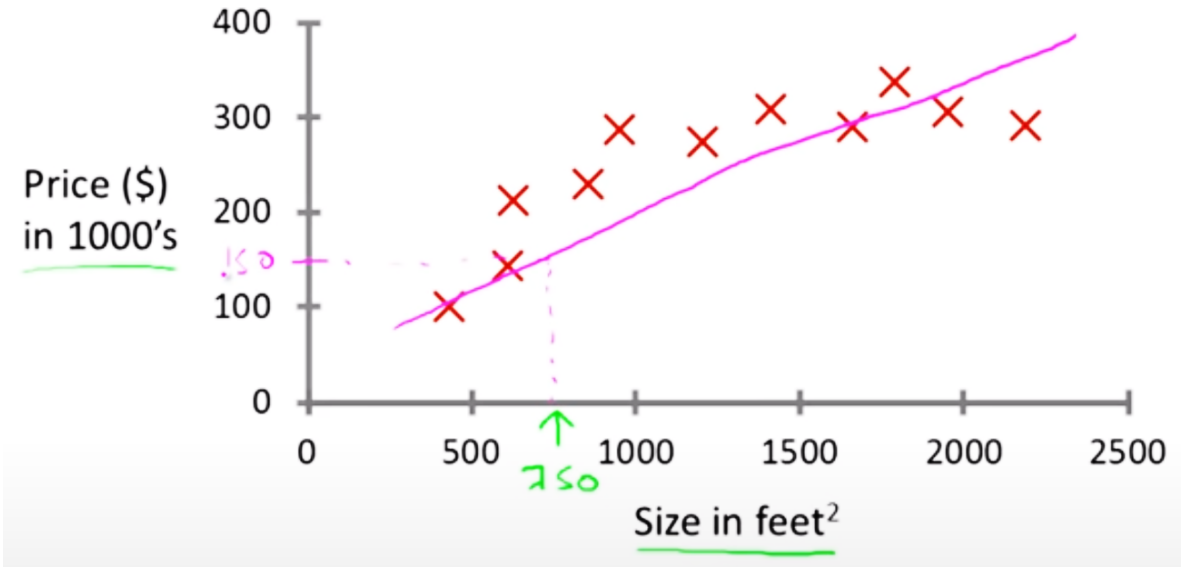
Predictive Model: Data Splitting



Predictive Model: Regression Performance

Error = Predicted – Actual

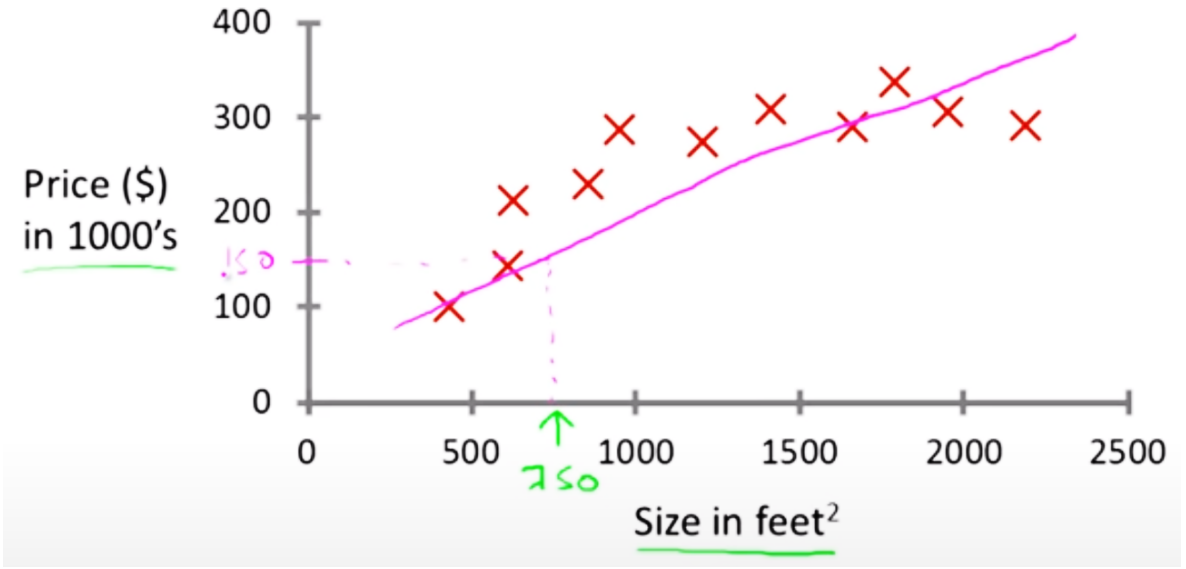
- To quantify the total amount of error, we take the square of each error and then sum all the squares, creating a measure called **Total Sum of Squared Error (Total SSE)**



Predictive Model: Regression Performance

Error = Predicted – Actual

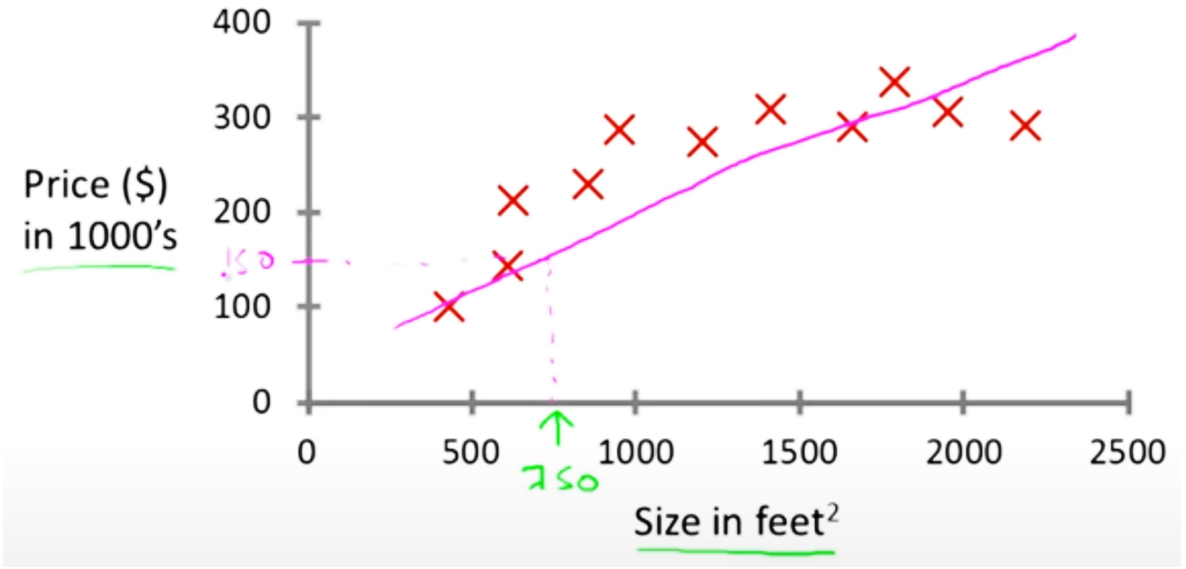
- Other popular measures that focus on error are **RMSE** (root mean squared error); **MAE** (mean absolute error); and **MAPE** (mean absolute percentage error)



Predictive Model: Regression Performance

Error = Predicted – Actual

- RMSE is useful because intuitively, it captures how much the predicted values diverge from the actual values on average



Predictive Model: Classification Performance

Accuracy = Number of correct classifications made / Total number of classifications

Predictive Model: Classification Performance

The **confusion matrix**

- Separates out the decisions made by the model, making explicit how one class is being confused for another

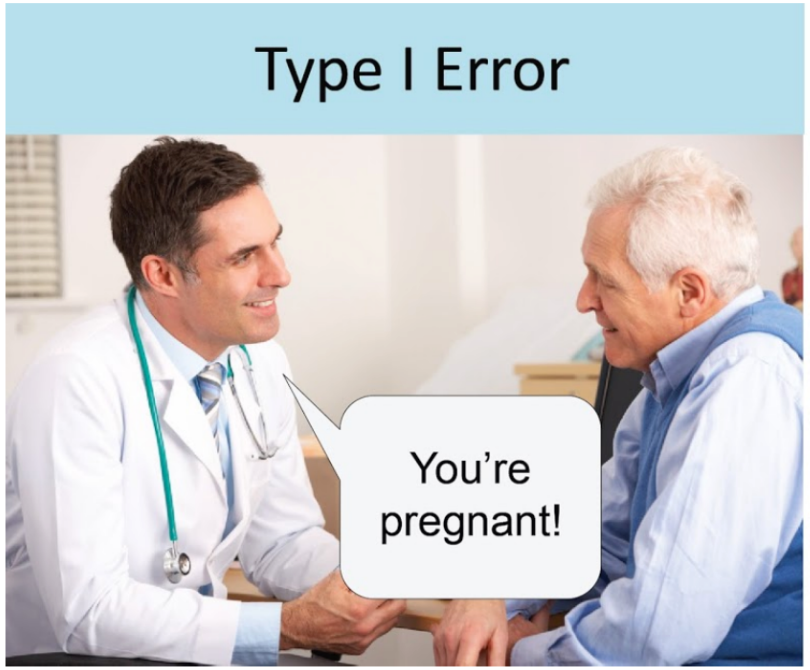
Type I Error

		Actual	
		Positive	**Negative**
Predicted	**Positive**	True Positives	**False Positives**
	Negative	**False Negatives**	True Negatives

Type II Error

Predictive Model: Classification Performance

False Positive



False Negative

Predictive Model: Classification Performance

The **F-Measure**

- Special-purpose combination of precision and recall
  - Precision = true positives / true positives + false positives
  - Recall = true positives / true positives + false negatives

- Always between 0 (worst performance) and 1 (best performance)

Things to Consider

- Is there a specific, quantifiable target that you are interested in predicting?
  - If yes, is it a **class or a number**?
    - Think about the decision

- Do you have data on the target?
  - Do you have **enough** data?
    - If the target is a class, a min of ~500 for each class type is needed

Another Thing to Consider

- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it

https://www.kaggle.com/alexisbcook/data-leakage#Example

Leakage

The use of data in the model training process that would not be available in practice at the time you would want to use the model to make a prediction, leading to **unrealistically good predictions**

- Leaking from the future into the past
- Leaking test data into the training data

Leakage (Example)

Suppose you want to predict who will get pneumonia. Below is a preview of your data:

got_penumonia	age	weight	male	took_antibiotic	…
false	65	100	false	false	…
false	72	130	true	false	…
true	58	100	false	true	…

People take antibiotics after getting pneumonia in order to recover. Using **took_antibiotic** to predict **got_pneumonia** will lead to data leakage

Avoiding Leakage

- Do you have relevant data **prior** to the decision?
  - Think about the timing of decision and action leading up to it

Summary

- A model is not reality, but rather a simplified version of it that serves a specific purpose
- Good models are useful, all models are wrong
- Predictive models estimate an unknown value, which can be a class (classification) or a number (regression)

Summary

- When building a predictive model, split data into training and testing sets (70-30 or 80-20)
- We evaluate classification models differently than regression models
- Many things to consider when building predictive models, including the data type of target variable, how much data are available, and issues of leakage

Thank You!