# What makes a good prediction?

K Arnold

# Objectives

- Compare and contrast regression tasks and classification tasks, and give examples of each
- Identify two different ways of measuring accuracy for regression and for classification
- Identify several reasons why a model may predict better on some subsets of data than others

# Types of Tasks

- **regression**: predict a *number* ("continuous")
  - number should be "close" in some sense to the correct number
- **classification**: predict a *category*
  - which one of these two groups? three groups? 500,000 groups?
  - could ask: "how likely is it to be in group *i*"

# Are these tasks *regression* or *classification*?

1. Is this a picture of the inside or outside of the restaurant?
2. How much will it rain in GR next year?
3. Is this person having a seizure?
4. How much will this home sell for?
5. How much time will this person spend watching this video?
6. How big a fruit will this plant produce?
7. Which word did this person mean to type?
8. Will this person "Like" this post?

# Today's examples

**Regression**: housing prices in Ames, Iowa. Details:

- Paper
- Data Dictionary

**Classification**: *seizure classification*.
First FDA-approved AI-powered medical device: Empatica Embrace2, company founded by MIT data scientist Rosalind Picard

# What makes a good prediction? *Regression*

We predicted the home would sell for $250k. It sold for $200k. Is that good?

# What makes a good prediction? *Regression*

We predicted the home would sell for $250k. It sold for $200k. Is that good?

- **residual**: actual minus predicted
  - If home sold for $200k but we predicted $250k, residual is _
- **absolute error**
- **squared error**

# What makes a good prediction? *Regression*

We predicted the home would sell for $250k. It sold for $200k. Is that good?

- **residual**: actual minus predicted
  - If home sold for $200k but we predicted $250k, residual is _
- **absolute error**
- **squared error**

Across the entire dataset:

- **average error**: do we tend to predict too high? too low? "*bias*"
- **max** absolute error
- **mean** absolute error
- **mean squared error** (MSE)
- normalized squared error: MSE / Variance
  - The confusingly named "R2" = 1 - normalized squared error

# What makes a good prediction? *Classification*

Suppose: every minute, the armband decides whether a seizure is occurring

The child was perfectly fine but our armband flagged a seizure. Is that good?

# What makes a good prediction? *Classification*

Suppose: every minute, the armband decides whether a seizure is occurring

The child was perfectly fine but our armband flagged a seizure. Is that good?

The child was having a seizure but our armband didn't flag it. Is that good?

# What makes a good prediction? *Classification*

| | **Seizure happened** | **No seizure happened** |
|---|---|---|
| Seizure predicted | True positive | False positive (Type 1 error) |
| No seizure predicted | False negative (Type 2 error) | True negative |

# What makes a good prediction? *Classification*

| | **Seizure happened** | **No seizure happened** |
|---|---|---|
| Seizure predicted | True positive | False positive (Type 1 error) |
| No seizure predicted | False negative (Type 2 error) | True negative |

- **Accuracy** (% correct) = (TP + TN) / (# episodes)
- **False negative** ("miss") **rate** = FN / (# actual seizures)
- **False positive** ("false alarm") **rate** = FP / (# true non-seizures)

# What makes a good prediction? *Classification*

| | Seizure happened | No seizure happened |
|---|---|---|
| Seizure predicted | True positive | False positive (Type 1 error) |
| No seizure predicted | False negative (Type 2 error) | True negative |

- **Accuracy** (% correct) = (TP + TN) / (# episodes)
- **False negative** ("miss") **rate** = FN / (# actual seizures)
- **False positive** ("false alarm") **rate** = FP / (# true non-seizures)
- **Sensitivity** ("true positive rate") = TP / (# actual seizures)
  - Sensitivity = 1 − False negative rate
- **Specificity** ("true negative rate") = TN / (# actual seizures)
  - Specificity = 1 − False positive rate
- Wikipedia article

If you were designing a seizure alert system, would you want sensitivity and specificity to be high or low? What are the trade-offs associated with each decision?

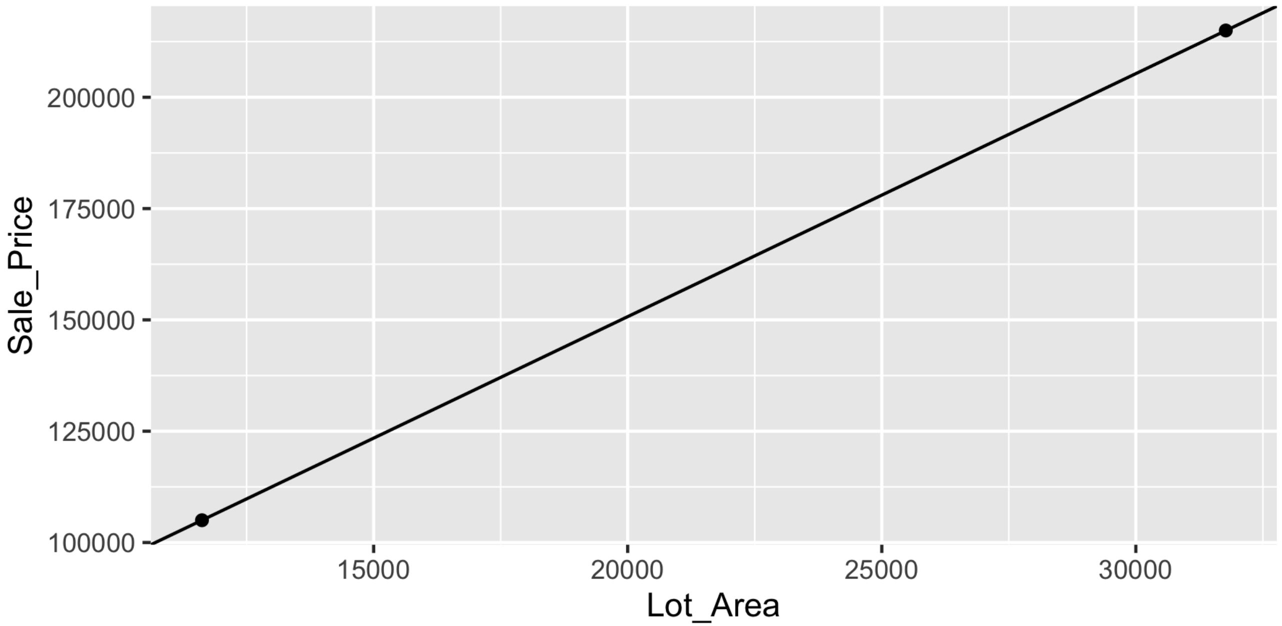
# Validation

**Key point**: you *must* evaluate predictions on *unseen* data

Hey look! I can exactly predict how much a home will sell for!

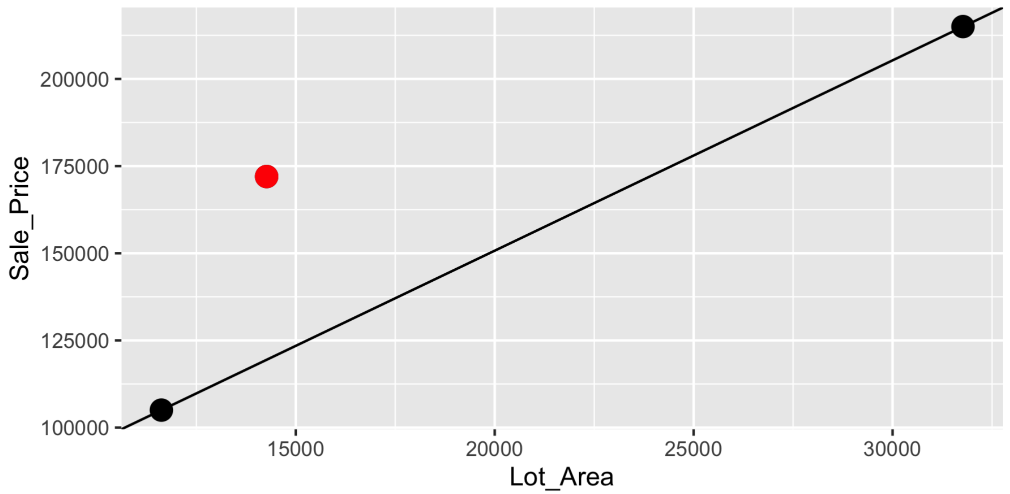
| Lot_Area | Sale_Price |
|---|---|
| 31770 | 215000 |
| 11622 | 105000 |

sale price = 41548.54 + 5.459599 * lot area



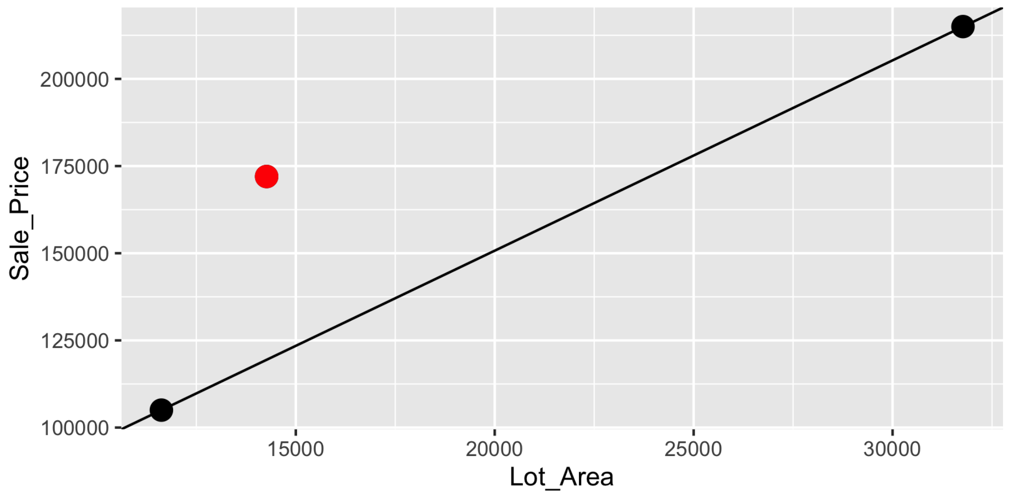
# Validation: *unseen* data

| Lot_Area | Sale_Price |
|---|---|
| 31770 | 215000 |
| 11622 | 105000 |
| 14267 | 172000 |



# Validation: *unseen* data

| Lot_Area | Sale_Price |
|---|---|
| 31770 | 215000 |
| 11622 | 105000 |
| 14267 | 172000 |




| Lot_Area | Sale_Price | predicted | residual |
|---|---|---|---|
| 31770 | 215000 | 215000.0 | 0.00 |
| 11622 | 105000 | 105000.0 | 0.00 |
| 14267 | 172000 | 119440.6 | 52559.36 |

# Oh ok, I'll just fix that one...

| Lot_Area | Bsmt_Unf_SF | Sale_Price |
|---|---|---|
| 31770 | 441 | 215000 |
| 11622 | 270 | 105000 |
| 14267 | 406 | 172000 |

sale price = -37769.46 + 1.5311432 * lot area + **462.8685748 * basement sq ft**

## and look, it works!

| Lot_Area | Bsmt_Unf_SF | Sale_Price | predicted | residual |
|---|---|---|---|---|
| 31770 | 441 | 215000 | 215000 | 0 |
| 11622 | 270 | 105000 | 105000 | 0 |
| 14267 | 406 | 172000 | 172000 | 0 |

*Do you really think so?*

# Failure to generalize

Predictive models almost always do better on the data they're trained on than anything else.
Why?

- model uses a pattern that only held by chance
- model uses a pattern that only holds for some data
- model uses a pattern that's real but got a fuzzy picture of it

General name: **Overfitting**