GENERAL ASSEMBLY

INTRODUCTION TO CLASSIFICATION

Matt Brems

General Assembly, DSI+

DATA SCIENCE PROCESS

1. Define problem.
2. Gather data.
3. Explore data.
4. Model with data.
5. Evaluate model.
6. Answer problem.

MODELING

- Modeling is something that we naturally do.
- A **model** is a simplification of reality.
- How do we simplify?
  - Taking into account only really important factors.
  - Making assumptions about how things behave.

MODELING TECHNIQUES

INTRODUCTION TO CLASSIFICATION

- When our *Y* variable is **continuous**, then we refer to this problem as a **regression problem**.
  - Examples:

- When our *Y* variable is **discrete**, then we refer to this problem as a **classification problem**.
  - Examples:

TURNING REGRESSION PROBLEMS INTO CLASSIFICATION PROBLEMS

- We can turn our continuous Y into a discrete Y.
  - Voting Example:
  - Lyft Example:

WHAT IS SO DIFFERENT ABOUT CLASSIFICATION?

- Our observed values may or may not be ordered.

- I will evaluate my model differently.

WHAT IS THE SAME ABOUT CLASSIFICATION & REGRESSION?

- The bias-variance trade-off is still **extremely** relevant!
  - We still (**always**) do a train-test split!
  - We can still cross-validate!
  - We can still regularize! (Remember to scale your features before regularizing.)
  - We can still do feature selection/engineering!

- We will sometimes still have hyperparameters (a.k.a. tuning parameters).

- We're still simplifying reality with a model… our Y just looks a little different.