DSND: Introduction To Data Science

->Lesson 1: The Data Sci. process

How to approach DS problems

-> CRISP-DM

cross industry standard process 4 data mining

1. Delop business understanding
   - ↳ understanding problems & pain points
2. Data understanding
   - ↳ what data will you need to solve your problem?

↳ note that 1 & 2 can happen in either order

3. Prepare data
   - ↳ Collect, access data
   - ↳ Wrangling & feature engineering where possible/needed

(4) Data Modeling

↳ Perform statistical analysis

↳ Answer questions

(5) Evaluate results

(6) Deploy / Iterate

<u>what is required</u>

- Curiosity
- the right data
- the right tools

<u>Modeling</u>

→ Instantiate, fit , predict, score

- Quantitative vs Categorical vars.

Looking at variables

↳ correlation matrices, heatmaps, matrix plots etc } Choose features

Objective: try to capture as much signal in your data as possible

→ Handling missing values

① Drop rows

↳ problem, increases bias

② Impute values

↳ removes variability & predicting power of that column

→ Categorical variables

↳ ML models need numerical values

→ one hot encoding

↳ each choice in a categorical variable becomes a dummy variable

↳ Always remember to drop one column

Problems: does not scale well w/ many categorical variables.

→ Deploying Model

↳ Automate tasks

↳ Communicate for others to take action

↳ Dashboards

↳ etc.

→ Lesson 2: Communication w/ Stakeholders

↳ value of a project can be cut short

→ Github

↳ Repository for code

- Covered by README:
  - Motivation
  - versions & installs
  - file desc.
  - Liscensing

→ Medium

↳ communicate project w/ other people

→ Communication

↳ Who is the audience?

↳ technical vs accessible

↳ consider prior knowledge

① "Pull in" reader

② keep engaged w/ story

③ end w/ summary & call to action

① "Pull in"

↳ compelling image boosts engagement

- revevance
    - to others
    - to current events

② keep engaged

- use paragraphs!
- pictures & other whitespace for pacing

③ call to action

- reiterate main points
- call to action
    - ↳ makes clear how reader should act afterward

→ Pick a dataset

→ Pose 3 questions you want answered

→ Analyze data using python

  ↳ Munge

  ↳ Visualize

  ↳ Model

  ↳ etc.

→ Communicate insights

Deliverable: - Github repo

+

- Medium Blog Post.

Notes:

- Stick to CRISP DM