Career Lab 2020

MAXX POTENTIAL

Part 1: Business Understanding

Let's explore what we can do with a cool data set. You'll each need to set your own objective for this activity.

The five types of questions answered with Data Science:

1. How much or how many? (regression)
2. Which category? (classification)
3. Which group? (clustering)
4. Is this weird? (anomaly detection)
5. Which option should be taken? (recommendation)

Recommendations for setting your objective:

- What fields are in your dataset?
- Why would any of those fields be recorded?
- What could each field tell us about the organization, industry, etc.?
- Are any of those fields related to each other?
- Is there a causal or correlation relationship?

Some examples to help frame your objective:

1. Given: list of food items, cost per serving of each, and various nutrient levels per serving of each
   Use: {to be defined later}
   To: Identify the lowest cost diet selection that meets the minimum daily nutrient levels?
2. Given: 20 years of daily high temperature data from July through October
   Use: {to be defined later}
   To: Identify the true end of summer based on temperature drop

How businesses talk about their data advancements:

- https://www.sas.com/en_us/customers.html
- https://www03.ibm.com/software/businesscasestudies/us/en/corp
- https://www.informs.org/Impact/O.R.-Analytics-Success-Stories

Part 2: Capture & Maintain

For this activity we will skip these steps by utilizing an existing data set. If you want to explore Capture & Maintain, MySQL, Beautiful Soup, and Google Analytics for Firebase are some pathways to get comfortable with Data Mining and Maintenance.

Great Data Sources:

- https://www.data.gov/
- https://www.kaggle.com/
- https://data.world/datasets/open-data
- Many others out there, feel free to find data with any other tool

Part 3: Data Cleaning

The most time consuming part of the process. Here are some common issues with data sets you'll want to look for and resolve:

- **Data Response Type!** Data types could be inconsistent within a column (ex. Recorded as 0 or 1 most years, but one year was recorded as no or yes)
- **Data Storage Type!** Data types could have been stored inconsistently (ex. Using integer 0 vs string "0")
- **Typos/Capitalization!** In categorical data, there could have been typos, or multiple forms of the same category (ex. JavaScript, JS, javascript typically all point to JavaScript; male, Male, m, M, Maale → Male)
- **Missing Values!** If a column is missing values for > 5% of the records, typically this is substantial enough to not be able to rely on that value and there isn't an easy solution. If one or multiple columns each have < 5% of their values missing, we can proceed with a couple options
    - Ignore/throw out the records that are missing values. Depending on the size of the data set, this could be an unrealistic choice
    - For numerical data: Mean Imputation/Average Imputation → find the average value for the column, and fill the missing values with that average. Watch for bias though! Example, if all of the rows with a missing value for age happen to be all classified as female, we could accidentally add bias
    - For categorical data: Mode Imputation → similar to Mean Imputation, find the most common category value in the column, and fill the missing values with that value. Same caveat applies that we need to watch out for bias

- ADVANCED: Linear Regression Imputation → If you're already familiar with modeling, it is possible to build a model to predict what those values might be based on other values in the record.

You can work to resolve these issues by hand using a standard spreadsheet tool (Excel, Google Sheets, Numbers, etc.)

Challenge Route: Resolving these issues can be a lot faster using Python and Pandas

Part 4: Data Exploration

The fun brainstorming! With a newly cleaned data set, its time to re-evaluate the initial business case Given/Use/To statement. You might be able to formulate some initial hypotheses, and use some exploratory visualizations to help confirm you're tackling a realistic problem.

There are so many techniques and tools out there, this resource gives a good overview of what you might want to do next: https://en.wikipedia.org/wiki/Exploratory_data_analysis

Part 5: Feature Engineering

This is one of the sections that requires some advanced knowledge to determine what features to use, how to construct your various train/test/validate data sets, etc.

While we're skipping this section for now, Introduction to Feature Selection Methods, Feature Selection with sklearn, and Best Practices for Feature Engineering are some great resources to learn more about what this process would look like.

Part 6: Predictive Modeling

Again, to implement some Machine Learning models, you'll need some understanding of the models and how they can be utilized to solve problems. If you want to dive deeper, Udacity, Essentials of ML Algorithms, and Evaluating ML Models are great starting points.

Part 7: Data Visualization

Let's communicate what you've learned so far from your data set! Here are some resources for getting started with each of the various tools we've mentioned

- Excel
- Google Sheets
- Numbers
- seaborn
- Bokeh
- plotly