Scout (Mugdha) Sujyot

Week 4 paper

What are the top 3 ideas you retained from the material read in chapter 4?

1. It's very important to know how to handle missing values in your data. What you do with missing values depends on how many there are, and whether they're missing randomly or systematically. You can drop the rows with the missing values (but you should make sure they're not a large portion of your data) or convert them to a meaningful value (like a "missing" flag in categorical data for example).

1. There are other transformations that you can use to address issues and make the data easier to model and understand. Some useful ones are converting continuous variables to discrete, normalising variables, and log transformations. Normalisation and rescaling are important when relative changes are more important than absolute ones. Normalising by mean and standard deviation is most meaningful when the data is roughly symmetric. You can use log transformations for skewed and wide distributions to restore symmetry.

3. Data provenance records help reduce errors because data science is an iterative process. This helps you keep track of your data management steps as the data and models evolve.