Working through am analytic problem

--Hands on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems

--Adjusted for HCAN361 🡪 removed items related to machine learning

1. Frame the Problem and Look at the Big Picture
   1. Understand or define the objective
   2. Who are the consumers of your solution?
   3. “Conceptually,” what are the key measures?
   4. What are comparable problems? Can you reuse what has been done before?
   5. Is domain knowledge outside of your team needed?
   6. Identify potential gaps in data where assumptions are needed to come to a solution.
   7. List assumptions you have made and those provided by stakeholder.
   8. Verify your assumptions with stakeholder, and reaffirm project
   9. Establish preliminary timeline
2. Get the Data
   1. List the (both internally and externally sourced) data needed
   2. Find data and document information about the data
   3. Understand storage needs and create a workspace
   4. Map conceptual key measures to data
   5. Get data
   6. Using code copy needed data to an accessible digital workspace
   7. Ensure sensitive data is removed or protected
   8. Reassess project timeline
3. Explore the Data
   1. Study each attribute (i.e., key measure) and its characteristics
      1. Column name
      2. Data type (e.g., categorical, floating, integer, string, etc.)
      3. Percent of missing values
      4. Examine distributions for outliers (numeric), rare occurrences (string), and type (e.g., normal, left skewed, right skewed, uniform, etc.)
      5. Rank and re-rank measure’s usefulness to project relative to key measure
   2. Visualize data
   3. Study correlations between measures
   4. Identify potential data transformations (e.g., log transform skewed interval data, reduce categories in categorical variable, convert continuous interval data to groups, etc.)
   5. Assure that all data gaps/assumptions are addressed
   6. “Plan” the analysis
   7. Document progress
   8. Run by project supervisor/stakeholder to keep them updated
      1. This may be in the form of documentation accessible to stakeholders
4. Prepare the Data
   1. Make a “copy” of the “copy” (i.e., make a copy of data generated in “get the data”)
   2. Perform data cleaning, and other data transformations
   3. Remove or fill in rows with missing values
      1. If missing data ≤ 10% and “missing at random,” you can likely remove rows with missing observations
      2. If missing data > 10% and ≤ 30% and “missing at random,” then you can fill in missing data
      3. If missing data > 30%, then column may not be useful in analyses
   4. Perform feature selection (i.e., drop attributes/columns that are not useful)
5. Short-List of Promising Models
   1. Look at several ways to address objective
   2. For each model/algorithm, analyze and compare performance
   3. Select model with best performance and least amount of biases
6. Fine-Tune Selected Model
   1. Clean code and add comments to make code easier to read. Should be good enough to hand-off and allow another analyst to independently verify results
   2. Document the project’s methodology, results, and conclusion(s)
7. Present Your Solution
   1. Very important to highlight big picture at beginning and tie your findings back to big picture at end of presentation
   2. Make sure to appropriate time to discuss why presented solution achieves the designated objective(s)
   3. Be sure to highlight key assumptions and limitations of the analytic project
   4. Do not feel compelled to discuss “every aspect” of project. Reference the project’s documentation so that the audience can know where all details are located. If the audience wants to know something you did not mention, they will ask.
   5. When possible, ensure presentation displays beautiful visualizations and easy-to-remember statements.