Exercise 7a: Your Problem, Formulated as an ML Problem

Write down or click on the checkbox for what you think is the best technical solution for your problem. Our problem is best framed as:

- [x] Binary classification
- [ ] Unidimensional regression
- [ ] Multi-class single-label classification
- [ ] Multi-class multi-label classification
- [ ] Multidimensional regression
- [ ] Clustering (unsupervised)
- [ ] Other

which predicts:

the amount of time that has passed between receipt of email and sending the first response.

8 - Data for model
Time in unopened state, Time passed between receipt and response, Starred by user, Count of important keywords in title, Label = Important

Real time - Time in unopened state, time passed between receipt and response, starred by user
Pre-process pipeline - count of important words

Exercise 7b: Cast your Problem as a Simpler Problem

When first starting out, simpler problem formulations are easier to reason about and implement. Take your given problem and state it as a binary classification or a unidimensional regression problem (or both).

Exercise 8: Design your Data for the Model

Write the data you want the ML model to use to make predictions.

Input 1	Input 2	Input 3			Output (label)

Tips for Success

- One row constitutes one piece of data for which one prediction is made.
- Only include information that is available at the moment the prediction is made.
- Each input can be a scalar or 1D list of integer, float, or bytes (including strings).
- If an input has a structure different from a scalar or 1D list, you may wish to consider whether that is the best representation for your data. For example:
  - If a cell represents two or more semantically different things in a 1D list, you may wish to split these into separate inputs.
  - If a cell represents a nested protocol buffer, you may wish to flatten out each field of the nested protocol buffer.
  - Exceptions: audio, image and video data, where a cell is a blob of bytes.

Exercise 9: Where the Data Comes From

Write down where each input comes from. Assess how much work it will be to develop a data pipeline to construct each column for a row.

Input 1	Input 2	Input 3			Output

Tips for Success

When does the example output become available for training purposes?

- If the example output is difficult to obtain, you might want to revisit Exercise 5 (Using the Output), and examine whether you can utilize a different output for your model.

Make sure all your inputs (except the output) are available at serving time (when the prediction is made), in exactly the format you are writing down.

- If it is difficult to obtain all your inputs at serving time in exactly the same format, you may want to revisit Exercise 8 (Design your data for the model) to reconsider inputs, or Exercise 5 to reconsider when serving can be made.

Exercise 10: Easily Obtained Inputs

Among the inputs you listed in Exercise 8, pick 1-3 inputs that are easy to obtain, and that you believe would produce a reasonable, initial outcome.

Input 1	Input 2	Input 3			Output

Tips for Success

- In Exercise 6, you listed a set of heuristics you could use. Which inputs would be useful for implementing these heuristics?
- Consider the engineering cost to develop a data pipeline to prepare the inputs, and the expected benefit of having each input in the model.
- Focus on inputs that can be obtained from a single system with a simple pipeline. Starting with the minimum possible infrastructure is advisable when first starting out.

Click the button below to either print or save your responses as a .pdf.

Print Page