

Source:#ret

0.1 | prompt:

For each of the scenarios below, answer the following questions. You do not have to explain your answers.

1. What type of machine learning problem (regression, classification, clustering) do you think this is?
2. If this is a supervised problem, what will you use as your targets (aka labels) and how will you get them?
3. What processing do you need to do to your input data? (How will you handle non-numerical inputs? Do you need to tokenize? etc.)
4. What type(s) of model(s) would you try? Remember to start with the simplest thing that might work! Try to justify your choice.
5. What validation metric(s) would you use to decide how well you're doing?
6. What ethical challenges do the data collection, creation, and/or use of this model create? If you feel there are none, explain why.

Scenarios:

1. You are playing fantasy football and want to predict how many points each player will score next season based on their performance this season.
2. You have customer reviews, each one of which has a rating from 1 (worst) to 10 (best) and some text. You want to predict the rating from the text.
3. You have data from a movie streaming service that consists of lists of movies that each user has watched and rated. You want to predict the rating from the list of movies.
4. You want to predict whether a random stranger owns a cat, a dog, or neither, based on things that they have posted on social media.
5. You want a model to predict the number of deer that will be born in a breeding season. You have a list of features that you think might be relevant:
 - number of fawns born
 - the genus and species
 - number of does sighted during the mating season
 - vegetation quality during the mating season ("low", "average", or "high")

0.2 | Scenarios:**1. Football**

1. Since we only have one season of point values, and hence cannot see cross season change in point values, the old season players will be used as training data.
2. Regression
3. Label: Point value
4. One Hot Encoding, 0-1 normalization
5. Linear Regression or Neural Networks
6. RMSE
7. None

2. Customer Reviews

1. Classification
2. Positive, Negative, Neutral
3. Tokenization,