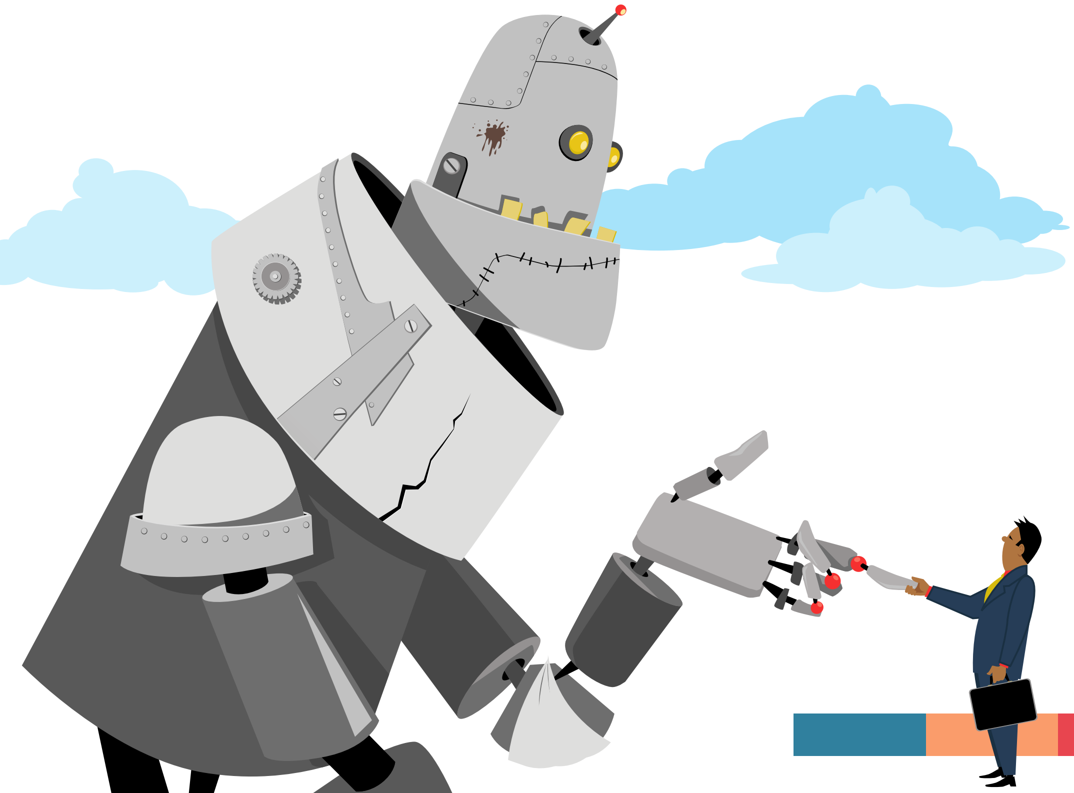
COMPANION WORKBOOK



DATA SCIENCE PRIMER

by EliteDataScience

CHAPTER | 01

BIRD'S EYE VIEW

Visit the online lesson.

***Note:** To save your inputs, you must use an offline PDF viewer such as Preview (Mac) or Acrobat (PC).

What are the 5 core steps of the machine learning workflow?

When the curious child learned that "red and bright means pain," what did he learn?

○ (A) An algorithm.
○ (B) A pattern.
○ (C) A model.
○ (D) Both (B) and (C).
○ (E) None of the above.

In the example of the curious child, what was the training data? What was the test data?

In your own words, describe the 3 essential elements of great machine learning.

CHAPTER | 02

EXPLORATORY ANALYSIS

Visit the online lesson.

What types of features can have sparse classes? How would you check for them?

What does it mean if ‘sqft’ (size of property) has a correlation of 0.68 with ‘baths’ (# of bathrooms)?

What are 3 sanity checks to make by looking at example observations from the dataset?

CHAPTER | 03

DATA CLEANING

Visit the online lesson.

What are 2 types of unwanted observations to remove from the start?

What are 3 types of structural errors to look out for?

How should you handle missing data?

Why is it sub-optimal to drop observations with missing data or impute missing values?

CHAPTER | 04

FEATURE ENGINEERING

Visit the online lesson.

What are indicator variables and why are they useful?

What are two criteria you can use to group sparse classes?

In a set of dummy variables created from the same feature, would there ever be multiple variables with value 1 (per observation)?

In our real-estate example, what would be the values for the ‘exterior_walls’ dummy variables if a property had metal walls?

CHAPTER | 05

ALGORITHM SELECTION

Visit the online lesson.

What are the two biggest flaws of linear regression?

How can you address the first flaw? (Which mechanism, and which algorithms?)

How can you address the second flaw? (Which mechanism, and which algorithms?)

What are two types of regularization penalty, and what do they do in practice?

What are two methods for ensembling and how do they work?

CHAPTER | 06

MODEL TRAINING

Visit the online lesson.

(Hopefully a freebie) Pick one: better data or fancier algorithms.

When should you split your dataset into training and test sets, and why?

What’s the key difference between model parameters and hyperparameters?

Explain how cross-validation helps you “tune” your models?

CHAPTER | 07

NEXT STEPS

After completing this primer, these are the steps we recommend taking. You can read more detail about them here.

1.) Learn Python
Especially data science specific concepts like programming basics, data structures, flow control and functions, NumPy, and Pandas.

2.) Clarify Essential Theory
Especially concepts such as model complexity, mapping functions, causes of overfitting, cost functions, and machine learning algorithms.

3.) Master Core Skills
Specifically: exploratory analysis, data cleaning, feature engineering, algorithm selection, and model training.

4.) Build Situational Skills
These include data wrangling, preprocessing, and how to package your model into a script that can be run on the cloud.

5.) Practice Making Decisions
When have you done enough exploratory analysis? When should you pre-process your features? Which performance metrics should you use? And so on...

6.) Develop Advanced Skills
Skills including multi-step pipelines, handling the curse of dimensionality, PCA, probability thresholds, ROC curves, multi-layered groupbys, and advanced visualizations.

7.) Reinforce Key Concepts
Circle back and review all that you've learned, and then repeat the entire process with a few more end-to-end projects.

This is actually the exact process we've taken thousands of successful students through in our popular Machine Learning Masterclass. Check it out if you'd like our over-the-shoulder mentorship through this journey.

ANSWER KEY

CHAPTER | 01

BIRD'S EYE VIEW

Answer Key

What are the 5 core steps of the machine learning workflow?

1. Exploratory Analysis
2. Data Cleaning
3. Feature Engineering
4. Algorithm Selection
5. Model Training

When the curious child learned that "red and bright means pain," what did he learn?

(D) He learned both a pattern and a model.

"**Red and bright** means pain" is a pattern, and it became his model for dealing with bright, red objects.

Presumably, he could continue adding to that model. For example, a red and bright toy car would form a different pattern (and **heat** may become the distinguishing factor).

In the example of the curious child, what was the training data? What was the test data?

The training data was the candle flame.

The test data was the stove top. However, in this situation, we typically refer to the stove top as "unseen data."

In your own words, describe the 3 essential elements of great machine learning.

The first element is a "skilled chef." You must make dozens of decisions along the way.

The second element is "fresh ingredients." The quality of your data determines the effectiveness of your models.

The third element is to "avoid overcooking it." Overfitting a serious pitfall, and you must take precautions.

CHAPTER | 02

EXPLORATORY ANALYSIS

Answer Key

What types of features can have sparse classes? How would you check for them?

Categorical features can have sparse classes, and you can check for them using bar plots.

What does it mean if ‘sqft’ (size of property) has a correlation of 0.68 with ‘baths’ (# of bathrooms)?

It means that ‘sqft’ and ‘baths’ have a fairly strong positive correlation.

In other words, larger properties have more bathrooms, which makes sense.

What are 3 sanity checks to make by looking at example observations from the dataset?

Do the columns make sense?

Do the values in those columns make sense?

Are the values on the right scale?

Is missing data going to be a big problem based on a quick eyeball test?

CHAPTER | 03

DATA CLEANING

Answer Key

What are 2 types of unwanted observations to remove from the start?

Duplicate observations that can sometimes arise during data collection.

Irrelevant observations that don’t fit the specific problem you’re trying to solve.

What are 3 types of structural errors to look out for?

Typos

Inconsistent capitalization

Mislabeled classes

How should you handle missing data?

Missing categorical data should be labeled with a new class called ‘Missing’

Missing numeric data should be flagged (with a new indicator variable) and filled with 0.

Why is it sub-optimal to drop observations with missing data or impute missing values?

If you drop observations, you’re dropping information.

If you impute missing values, you’re obscuring the fact that the data was missing in the first place.

Remember, “missingness” is often informative.

CHAPTER | 04

FEATURE ENGINEERING

Answer Key

What are indicator variables and why are they useful?

Indicator variables are binary (0 or 1) variables that indicate if an observation meets a certain condition.

They allow you to isolate key properties, and they help you bring in your domain knowledge.

What are two criteria you can use to group sparse classes?

You can group similar classes.

You can also group the remaining sparse classes into a single 'Other' class.

In a set of dummy variables created from the same feature, would there ever be multiple variables with value 1 (per observation)?

No, because each dummy variable represents 1 possible class, and the original feature only had 1 class per observation.

In our real-estate example, what would be the values for the 'exterior_walls' dummy variables if a property had metal walls?

'exterior_walls_Metal' = 1, all others = 0

CHAPTER | 05

ALGORITHM SELECTION

Answer Key

What are the two biggest flaws of linear regression?

It’s prone to overfit with many input features.

It cannot easily express non-linear relationships.

How can you address the first flaw? (Which mechanism, and which algorithms?)

Regularization: Lasso, Ridge, Elastic-Net

How can you address the second flaw? (Which mechanism, and which algorithms?)

Decision tree ensembles: Random Forest, Boosted Tree

What are two types of regularization penalty, and what do they do in practice?

One type penalizes the absolute size of coefficients. This can cause coefficients to be zero, thereby performing automatic feature selection.

Another type penalizes the squared size of coefficients. This dampens coefficients, which is known as feature shrinkage.

What are two methods for ensembling and how do they work?

Bagging attempts to reduce the chance overfitting complex models. It combines many strong learners that were trained in parallel.

Boosting attempts to improve the predictive flexibility of simple models. It combines many weak learners that were trained sequentially.

CHAPTER | 06

MODEL TRAINING

Answer Key

(Hopefully a freebie) Pick one: better data or fancier algorithms.

Better data!

When should you split your dataset into training and test sets, and why?

You should split your dataset before you begin modeling, and you shouldn't touch your test set until you're ready to pick a final model.

This allows you to have a truly unseen dataset at the end, which will give you a reliable estimate of model performance.

What's the key difference between model parameters and hyperparameters?

Model parameters can be learned directly from the data (e.g. regression coefficients).

Hyperparameters must be set before training a model (e.g. penalty strength).

Explain how cross-validation helps you "tune" your models?

Cross-validation allows you to use only your training set to evaluate how an algorithm performs using different sets of hyperparameter values.

You can use the cross-validated scores (on the hold-out folds) to pick the best set of hyperparameter values.