Learning Bytes - Intro to Data Science for Finance

Rebecca Sealfon

Data Science

The field of extracting knowledge and insights from data.

May use mathematics, statistics, computer science and even the relevant social sciences and humanities as they relate to collecting and understanding the data.

"The Sexiest Job of the 21st Century" *-Harvard Business Review*

U.S. Bureau of Labor Statistics: 11.5 million data science job openings by 2026

Understanding the problem / Collecting the data

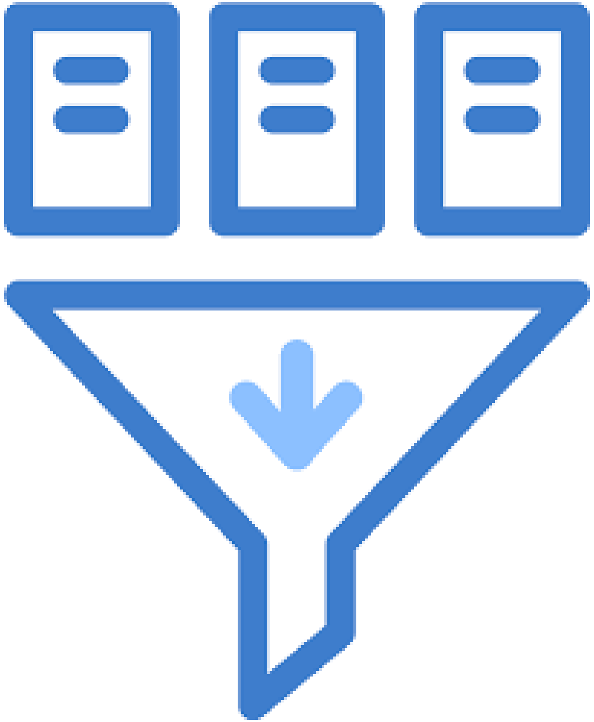
Understanding the problem

- E.g. talking to stakeholders, researching the topic
- Curiosity
- Domain-specific knowledge
- Methodicality
- Attention to detail
- Listening / observing
- Asking the right questions

Collecting the data

- After, during or before understanding the problem
- Web servers, logs, databases, application programming interfaces, online data repositories, etc.
- May require scraping, special integrations with data sources

Data preparation



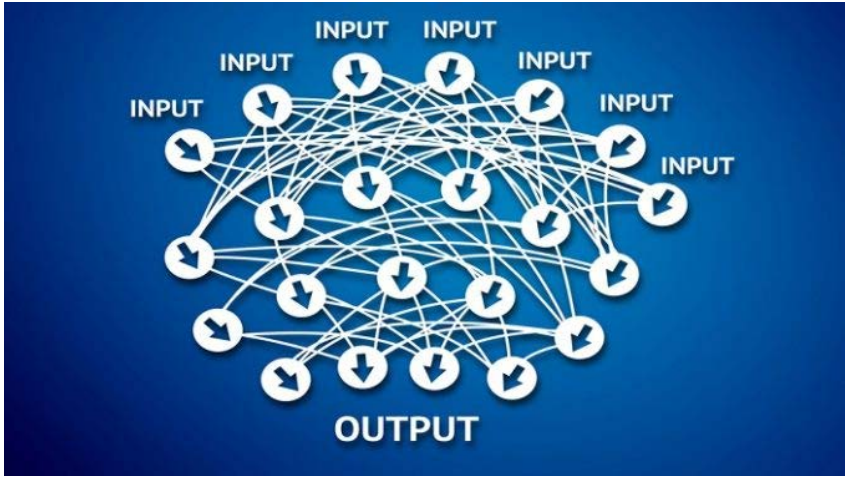
- Data cleaning
  - Most time-consuming step
  - Handle inconsistent datatypes, misspelled attributes, missing and duplicate values, etc.
- Data transformation
  - Modify data based on specific attributes

Exploratory data analysis

Defining, refining the set of inputs that will be used in model development

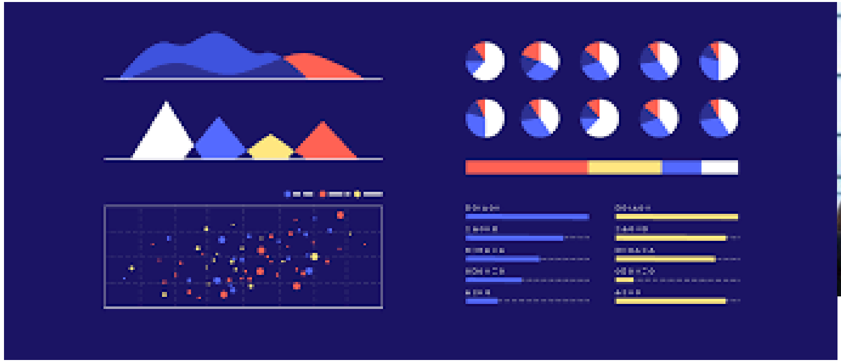
Required for an accurate model

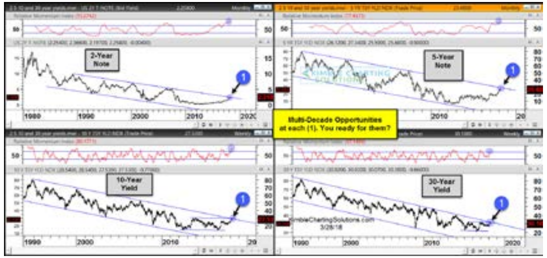
Data modeling

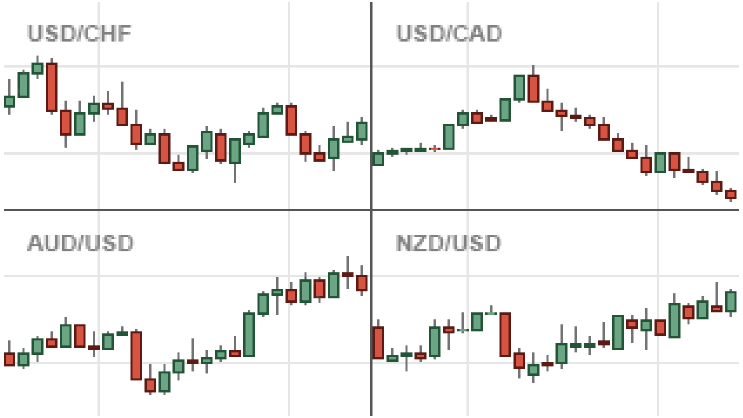


Repeated application of machine learning and other statistical techniques to identify the model that best fits the data and the requirements

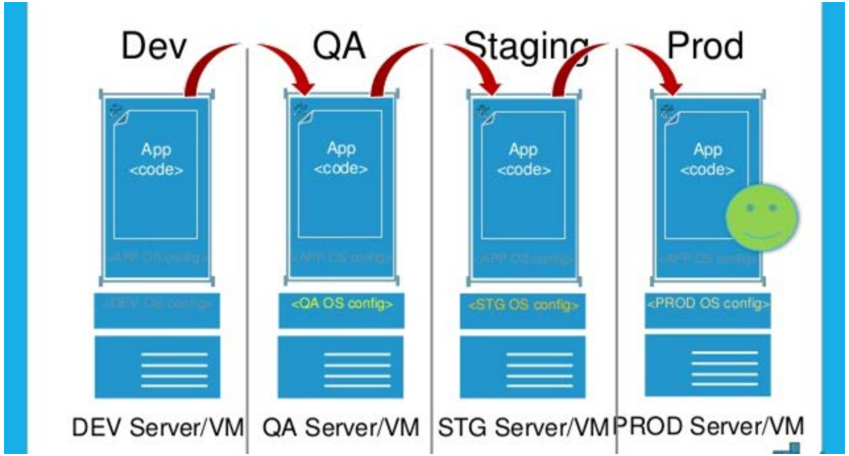
Data visualization and communication of findings







Deploying the model

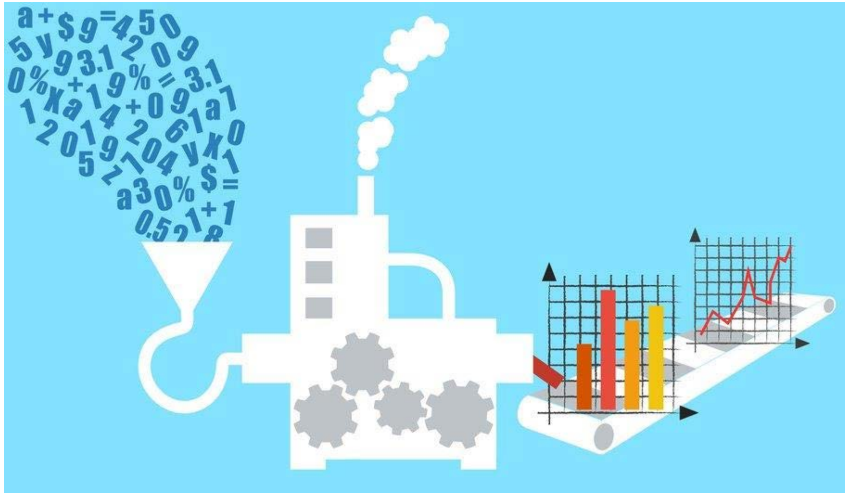


Typical sequence of deployment:

- Locally on one’s own computer
- DEV
- QA
- Production (what users see)

Industrial-grade deployment should follow standard sequence, regular schedule (e.g. DEV, QA, Production releases on specific days of week)

Maintaining the model



Ensure data analytics in place to continue to identify how to refine the model

Byte Academy Data Science Curriculum

- 14 weeks full-time, 24 weeks part-time at 2 evenings/week
- In-person, online options
- Data Science Immersive: 3 modules
  - Introduction to Programming: shared with Byte's Fullstack course
  - Data Science
  - Final Projects: industry-focused. Possible to collaborate with Byte Fullstack students
- Data Science Foundations: gentler introduction, can be pre-Immersive

- Curriculum designed to give students maximal flexibility in choosing, changing tracks
- 3:1 student to instructor ratio allows more feedback
- 90% job placement rate from Byte

Introduction to Programming Module

- Introduction to Python programming
- Navigating the computer terminal
- Building websites
- Working with shared code repositories such as GitHub
- **Data structures**: collections of data values, the relationships among them, and ways to manipulate the data
- **Algorithms**: systematic methods for solving problems based on a predetermined set of steps
- **Databases**: organized collections of data held in the computer

Data Science Module

- **Distributed processing**: The use of many computers to process large-scale data
- **Cloud computing**: automatic on-demand availability of computer system resources
- Basic through intermediate statistics
- Data visualization
- **Machine learning**: The use by computer systems of patterns and mathematical models to automatically make predictions about input data
- **Artificial intelligence**: Machines’ ability to solve problems
- **Natural language processing**: The field of programming computers to process language data

Final Projects

- What you want them to be!
- The instructors know you, your strengths and your interests by this point. So we can advise you about which careers might be the right fit for you, and how to position your project portfolio to grab the attention of the hiring managers.
- The most industry-focused section of the course

R vs. python™

- R is mainly used for statistical analysis
- R was built by statisticians
- R has excellent tools to communicate and visualize the results
- R has more libraries at present

- Python is a more general-purpose data science language
- Python code is easier to maintain than R
- Python is widely used for other purposes, such as Web development (e.g. Django framework)
- Python has a growing number of libraries

Your first taste of Python!

- Go to https://trinket.io/console, online Python console
- Type

```
>>> print(“Hello world”)
>>> print(“Hello Byte Academy”)
>>> dictionary = {‘Rose’: ‘Red’, ‘Violet’: ‘Blue’}
>>> print(dictionary[‘Violet’])
>>> dictionary[‘Violet’] = ‘Purple’
>>> print(dictionary[‘Violet’])
```

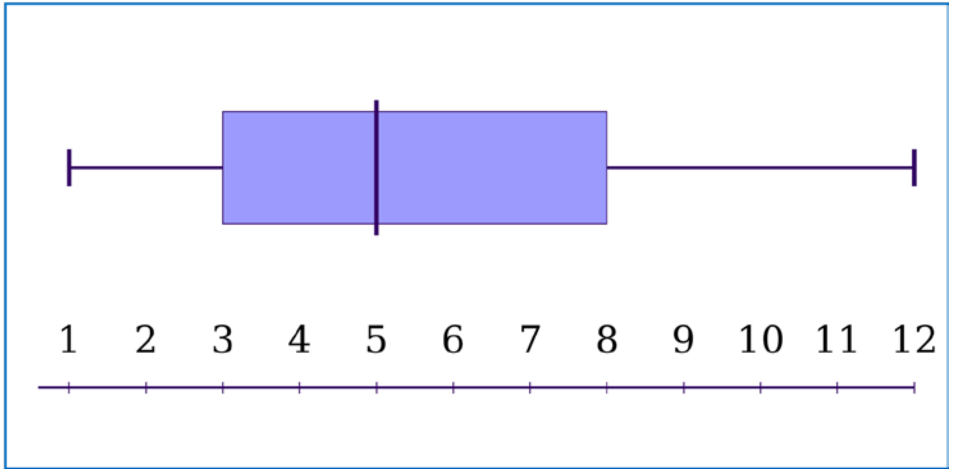
Data analysis and visualization

Length of words, viewed by quartile, in the definition of the word “pandas”

```
>>> from PyDictionary import PyDictionary as dictionary
>>> print(dictionary.meaning(“profit”))
{'Noun': ['the excess of revenues over outlays in a given period of time (including
depreciation and other non-cash expenses', 'the advantageous quality of being
beneficial'], 'Verb': ['derive a benefit from', 'make a profit; gain money or
materially']}
```

```
>>> words = dictionary.meaning("profit")['Noun'][0].split(" ")
>>> word_lengths = list(map(lambda x: len(x), words))
>>> print(word_lengths)
[3, 6, 2, 8, 4, 7, 2, 1, 5, 6, 2, 4, 10, 12, 3, 5, 8, 8]
>>> exit()
```

Go to https://www.meta-chart.com/box-and-whisker, copy and paste the word_lengths in the "Data" tab to visualize the data



Financial use cases

- Automating risk management
- Managing customer data
- Predictive analytics
- Real-time analytics
- Fraud detection
- Consumer analytics
- Algorithmic trading

Automating risk management

- Main steps: identifying, prioritizing, monitoring risks

Application: Identifying creditworthiness of potential customers. Done by credit report and other reports, references, and investigative methods, but machine learning continues to be useful to improve predictions of creditworthiness.

CREDIT REPORT

Your Credit Score is:

520

REJECTED

- Payment History: POOR
- Amount of Debt: POOR
- Length of Credit History: POOR
- Amount of New Credit: POOR

ICANT NAME: Mr. and Mrs. John Doe
RESS: 12345 Main Street
ATE: Anywhere
ZIP CODE:

Credit Report

Your Credit Score:

760

Credit Score Range:

Excellent

300 350 400 450 500 550 600 650

Excel
Good:
Fair:
Uncert
Poor:

Credit Score Grades:

Algorithmic trading

- Executing an order too large to fill at once, using automated pre-programmed instructions accounting for variables such as time, price, and volume to send small slices of the order (child orders) out to the market over time
- Machine Learning and other statistical models to improve trading algorithms based on results

Working with Big Data

- Go to http://www.kaggle.com
- Take a look around
- Which datasets do you find interesting? What other types of data might you want to see?

Data scientist salary

$95,000-$165,000

Further learning

- Python and Statistics for Financial Analysis, Coursera. Provided by Hong Kong University of Science and Technology. https://www.coursera.org/learn/python-statistics-financial-analysis
- Introduction to Data Science specialization, Coursera. https://www.coursera.org/specializations/introduction-data-science

Byte Academy
http://www.byteacademy.co
info@byteacademy.co
Twitter: @byteacademyco, @byteacademy