Introduction

Getting Started

“How do I get started?” — A question that I have received far too frequently over my last few years as a data scientist operating in the Biotechnology sector, and the answer to which never really seemed to have changed from person to person. My recommendation was generally along the lines of learning Python and data science through online courses, and following a few tutorials to get a sense of how things work. What I found was that the vast majority of scientists that I have encountered that are interested in learning data science and a number of its applications tend to always get overwhelmed by the large volume of resources and documentation available on the internet. From ‘Getting Started in Python’ courses to ‘Comprehensive Machine Learning’ guides, the vast majority of those who ask the question “How do I get started?” often find themselves confused and demotivated just a few days into their studies. This is especially true for scientist or researchers in the lab that do not usually interact with code, algorithms, or predictive models. Using the terminal command line for the first time can be unusual, uncomfortable, and to a certain extent, terrifying.

To address this, I decided to write this book. A one-stop-shop to give scientists, engineers and everyone in between a fast and efficient guide to getting started in data science. If you are not a coder, and do not intend to be, you have the option to read this book from cover to cover without ever using Python or any other resource. You will still manage to walk away with a strong foundation and understanding of machine learning, its capabilities, and what it can bring to the table within your team. If you are a coder, you have the option to follow along on your personal computer and replicate everything you see. All of the code shown in this book is inclusive, connected, and designed to be fully replicable on your own device. In addition, all of the code used in this book and the associated tutorials are all available online for your use. So, depending on what your expectations of the phrase “getting started” are, you will be able to use this book effectively and efficiently regardless of your intent to code. So how do we plan to get started?

Throughout this book, we will introduce topics and through a series of tutorials and guides catered to problems and use cases commonly observed and experienced in the Biotechnology sector. Unlike the many online course and tutorials available within the field, this book is connected together in the sense that materials are introduced in a chronological and *aufbau-like* manner. In less 250 pages, we will introduce the main concepts and ideas relating to Python, SQL, Machine Learning, Deep Learning, and Natural Language Processing. We will cover popular approaches, best practices, and important information every data scientist should know. In addition to all of this, we will also train and develop a number of predictive models that we will deploy to the cloud using Amazon Web Services (AWS), Heroku, and Python Anywhere (PA). Whether you plan to bring data science to your team, implement models yourself, or interview for a data scientist position, this book will equip you with the right tools and resources for your journey.

Software

Before we can get started, we will need to install a few things. The good news is that whether you are using a Mac, PC, or Linux, almost everything we will use is compatible for all platforms. There are three main items we will need to install: Python, Pip, MySQL, and DynamoDB. Let us go ahead and get started with Python. Python is what’s known as an interpreted language, meaning it does not need to be compiled as other languages do. For the purposes of this book, we will be using Python3.7. There are a number of ways you can install Python on your computer. You can install the language alone from Python.org. This will provide you with a python interpreter in its most basic form from which you can run commands and execute scripts. An alternative installation process that would also install pip, and a collection of other useful libraries can be done using the Anaconda distribution that can be retrieved from anaconda.com. To have a working version of Python and its associated libraries on your computer as quickly as possible, using Anaconda is highly recommended. Secondly, we will need to install libraries to assist in a few areas. Think of libraries as nicely-packaged portions of code that we can import and use as we see fit. Anaconda will, by default install a few important libraries, but there will always be others. We can install those on-the-go using pip. We will visit this in more detail in the next chapter. Finally, we will need a place to store and save all of our data throughout the analysis and preprocessing phases of our projects. For this we can use SQL, and DynamoDB. These are two very popular options amongst the data science community allowing users to store their data in both a relation and non-relational manner. Don’t worry about installing these now – we will have the installation process in the coming chapters.