Do A Data Science Project in 10 Days

Gangmin Li

2021-03-07

Contents

"Dont't touch it, It's hot!",
...

You now know it is hot. don't you?
How?

Because, you'v touched it!

Preface

Data science has gain popularity in recent years. Harvard Business Review called it the Sexiest Job of the 21st Century. I am not sure whether it is sexy or not but one thing is sure is that data science subjects are very popular among university students. This can be seen from the course selection. Data science-related courses like Big Data, Machine Learning, Statistics, and algorithms have huge student enrollment numbers.

Data science is a multidisciplinary field. It blends data mining, data analysis, statistics, algorithm development, machine learning, and advanced computing and software technology together in order to solve analytically complex problems. Its ultimate goal is to reveal insight into data and get the data value for the business. It is obvious that gaining the related knowledge is essential but my observation is that a lot of students shy away from doing any data science projects is because they are lacking hands-on experience in any full cycle of doing a data science project.

There is a Chinese saying that "Practice makes perfect". It is true but it is even more true that practices can gain the first hand of knowledge about practical issues and techniques to resolve the issues. Furthermore, it can build confidence in doing a data science project. That is what this book is intended to bring about.

What is This Book?

This book is originated from my Data Science course. It was in the lab sessions, where students practice different steps in the workflow of Data Science projects. Students enjoyed the detailed practices using different methods, algorithms, and techniques to solve analytical problems. To my surprise, we only find out later, that the student failed to grasp the real meaning of doing data science, which is NOT to provide an absolutely perfect working solution to a problem, rather, an experimental interpretation of data at hand. In other words, a data science project is normally aimed to provide an explanation of what the data is telling you. Even you are making a prediction model, there will never be a perfect

model that produces 100 percent prediction accuracy. You are not using data you have to solve problems that data has not to provide any solution to you!

I did a short tutorial for my students. The tutorial was emphasizing on the process and workflow of doing a data science project. That short tutorial was extremely successful and welcomed by all students, particularly the students who are not from Computer Science, Software Engineering, Statistics, Applied Mathematics, etc. rather, from Information Science and Management Science. I figured out the student's satisfaction comes from the practical skills and particularly the hands-on experience of doing a data science project rather than learning methods, algorithms, and parallel computation platforms without doing any.

So, I suppose this book is practical for students who have no background in computing and programming knowledge but interested in doing a Data Science project or moving to Data Science in the future.

In summary, this book is an introduction level book for novelty students who want to learn Data Science in a short period of time perhaps a few days or during their winter or summer holidays.

Structure of the Book

This book intended to follow the process of doing a typical data science project. That is the six steps of the data analytical process:

1. Understand the problem
2. Understand the data
3. Data Preprocess
4. Data analysis
5. Validation
6. Report

Each step, the tasks that need to be performed will be introduced and also practiced by an example project.

Before the example project to be kicked start, the tools and the platform are used in the example project will be introduced and practiced. The practice is basically mimicking the instructions on the book or the copy and paste code and run them in the environment.

What Can This Book Offer You?

This book offers a short practical course. When you finish the course, you will:

1. Understand the data analysis procedure
2. Familiar with R language
3. Using RStudio to do your prototype of data project
4. Data preprocess - manipulate data and get it ready for further analysis
5. Manage basic data analytical methods like **Descriptive Data Analysis**, **Exploratory Data Analysis** and **Predictive Data Analysis**
6. Have basic skills of visualize data results
7. Basic interpretation of your analyzing results
8. Report and communicate your results
9. Enter the data science community

Notes

Our goal is not to teach you R, but to teach you the basic process of doing a Data Science project that many other programming languages like Java and Python can do. We use R in our lessons because:

- we have to use something for examples;
- it's free, well-documented, and runs almost everywhere;
- it has a large (and growing) user base among scientists; and
- it has a large library of external packages available for performing diverse tasks.

But the two most important things are to use whatever language your colleagues are using, so you can share your work with them easily, and to use that language well. apparently. R is the most used language in Data Science.

Acknowledgements

During the process of writing this book, I have gained tremendous inspiration from many materials including but not exhausted the following, for which I owe them great gratitude.

1. A beginner's guide to Kaggle's Titanic problem Sumit Mukhija. (`https://towardsdatascience.com/a-beginners-guide-to-kaggle-s-titanic-problem-3193cb56f6ca`)
2. Kaggle competition. `https://www.kaggle.com/ldfreeman3/a-data-science-framework-to-achieve-99-accuracy`
3. Machine Learning for Dummies by John Mueller and Luca Massaron - Easy to understand for a beginner book, but detailed to actually learn the fundamentals of the topic

4. Data camp tutorials. `https://www.datacamp.com/community/`