***Data Science Using Python and R,***

**by Chantal Larose and Daniel Larose**

**Solutions Manual**

**Chapter 1: Introduction to Data Science**

**CLARIFYING THE CONCEPTS**

Solutions by Daniel Larose

1. **What is data science?**

Simply put, *data science* is the systematic analysis of data within a scientific framework. That is, data science is the adaptive, iterative, and phased approach to the analysis of data, performed within a systematic framework, that uncovers optimal models by assessing and accounting for the true costs of prediction errors.

1. **Which areas of study does data science combine?**

Data science combines the data-driven approach of statistical data analysis, the computational power and programming acumen of computer science, and domain-specific business intelligence,

in order to uncover actionable and profitable nuggets of information from large databases.

1. **What is the goal of data science?**

Data science lets people leverage large amounts of data and computing power to tackle complex questions. Patterns can arise out of data which could not have been uncovered otherwise. These discoveries can lead to powerful results, such as more effective treatment of medical patients or more profits for a company.

1. **Name the seven phases of the Data Science Methodology.**
2. Problem Understanding Phase
3. Data Preparation Phase
4. Exploratory Data Analysis Phase
5. Setup Phase
6. Modeling Phase
7. Evaluation Phase
8. Deployment Phase
9. **Why is it a good idea to have a Problem Understanding Phase?**

So that teams that have worked hard to solve a problem do not later find out that they solved the wrong problem. To get different teams on the same page regarding the data science project.

1. **Why do we need a Data Preparation Phase? Name three issues that are handled in this phase.**

Raw data from data repositories is seldom ready for the algorithms straight out of the box. Instead, it needs to be cleaned, or “prepared for analysis”. When analysts first examine the data, they uncover the inevitable problems with data quality that always seem to occur. Three issues that are handled in this phase are:

* 1. Identifying outliers and determining what to do about them.
  2. Transforming and standardizing the data.
  3. Reclassifying categorical variables.

1. **In which phase does the data analyst begin to explore the data to learn some simple information?**

The Exploratory Data Analysis Phase.

1. **Explain in your own words why we need to establish baseline performance for our models. Which phase does this occur in?**

We need to establish baseline performance for our models because our models need to outperform null models that, for example, always predict negative, or always predict positive.

1. **Which phase represents the heart of your data scientific investigation? Why might we apply more than one algorithm to solve a problem?**

The Modeling Phase represents the heart of our data scientific investigation. A particular problem may be amenable to more than one algorithm. In such a case we should apply all appropriate algorithms, and perform model selection to select the best performing model.

1. **How do we determine whether our predictions are any good? During which phase does this occur?**

Our models need to be evaluated against the baseline performance measures from the Setup Phase, in order to determine whether our predictions are any good. This occurs during the Evaluation Phase.

1. **True or false: The data scientist’s work is done with the Evaluation Phase. Explain.**

False. The data scientist should help with the deployment of the model in the Deployment Phase.

1. **Explain how the DSM is adaptive.**

By *adaptive*, we mean that sometimes it is necessary to return to a previous phase for further work, based on some knowledge gained in the current phase.

1. **Describe how the DSM is iterative.**

Sometimes we may use our experience of building an effective model on a similar problem. That is, the model we created serves as an input to the investigation of a related problem.

1. **List the most common data science tasks.**

Description, Estimation, Classification, Clustering, Prediction, and Association

1. **Which of these tasks have many non-specialists been doing all along?**

The description task is in widespread use around the world by specialists and non-specialists alike.

1. **What is estimation? In estimation, what must be true of the target variable?**

Estimation refers to the approximation of the value of a numeric target variable using a collection of predictor variables. Estimation requires that the target variable be numeric.

1. **What is the most widespread task in data science? For this task, what must be true of the target variable?**

Classification represents perhaps the most widespread task in data science, and the most profitable. The target variable must be categorical.

1. **What are cluster profiles?**

Detailed descriptions of the characteristics of each cluster.

1. **True or false: Prediction can only be used for categorical target variables. Explain.**

False. One may predict the stock market price (continuous target) of a stock three months in advance.

1. **For an association rule, what do we mean by support?**

The support refers to the proportion of records the association rule applies to.