CSC8631 Assignment Report

D Walmsley, C1053068

15/11/2021

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

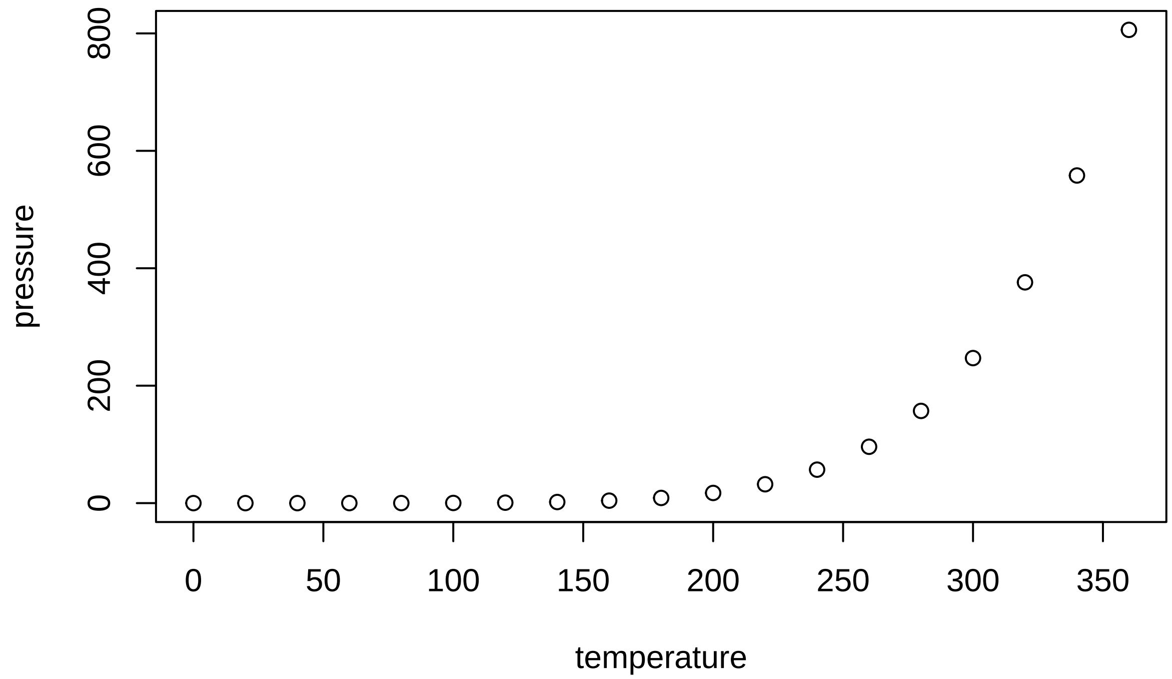
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Introduction

The purpose of this report is to analyse data collected from an online course ran by Futurelearn (www.futurelearn.com) that is ran in collaboration with Newcastle University. The report will demonstrate findings from a data set across 7 runs provided by the the university. Using the CRISP-DM methodology to provide a clear structure to the analysis. This report will be broken into sections that are the key steps of the CRISP DM process. The sections are *Business Understanding, Data Understanding, Data preparation, Modelling, Evaluation and Development.*

I.Business Understanding

The Business Understanding phase focuses on understanding the objectives and requirements of the project. Aside from the third task, the three other tasks in this phase are foundation project management activities that are universal to most projects:

Determine business objectives:

You should first "thoroughly understand, from a business perspective, what the customer really wants to accomplish." (CRISP-DM Guide) and then define business success criteria.

Assess situation:

Determine resources availability, project requirements, assess risks and contingencies, and conduct a cost-benefit analysis.

Determine data mining goals:

In addition to defining the business objectives, you should also define what success looks like from a technical data mining perspective.

Produce project plan:

Select technologies and tools and define detailed plans for each project phase.

II. Data Understanding

Next is the Data Understanding phase. Adding to the foundation of Business Understanding, it drives the focus to identify, collect, and analyze the data sets that can help you accomplish the project goals. This phase also has four tasks:

At first glance the data provided from the university is taken over seven years of running the course or runs. Data provided from runs one and two consist of files giving data on enrolments,

Collect initial data:

Acquire the necessary data and (if necessary) load it into your analysis tool.

Describe data:

Examine the data and document its surface properties like data format, number of records, or field identities.

Explore data:

Dig deeper into the data. Query it, visualize it, and identify relationships among the data.

Verify data quality:

How clean/dirty is the data? Document any quality issues.

III. Data Preparation

A common rule of thumb is that 80% of the project is data preparation.

This phase, which is often referred to as “data munging”, prepares the final data set(s) for modeling. It has five tasks:

Select data: Determine which data sets will be used and document reasons for inclusion/exclusion.

Clean data:

Often this is the lengthiest task. Without it, you'll likely fall victim to garbage-in, garbage-out. A common practice during this task is to correct, impute, or remove erroneous values.

Construct data:

Derive new attributes that will be helpful. For example, derive someone's body mass index from height and weight fields.

Integrate data:

Create new data sets by combining data from multiple sources.

Format data:

Re-format data as necessary. For example, you might convert string values that store numbers to numeric values so that you can perform mathematical operations.

IV. Modeling

What is widely regarded as data science's most exciting work is also often the shortest phase of the project.

Here you'll likely build and assess various models based on several different modeling techniques. This phase has four tasks:

Select modeling techniques:

Determine which algorithms to try (e.g. regression, neural net).

Generate test design:

Pending your modeling approach, you might need to split the data into training, test, and validation sets.

Build model:

As glamorous as this might sound, this might just be executing a few lines of code like "reg = LinearRegression().fit(X, y)".

Assess model:

Generally, multiple models are competing against each other, and the data scientist needs to interpret the model results based on domain knowledge, the pre-defined success criteria, and the test design.

Although the CRISP-DM guide suggests to "iterate model building and assessment until you strongly believe that you have found the best model(s)", in practice teams should continue iterating until they find a "good enough" model, proceed through the CRISP-DM lifecycle, then further improve the model in future iterations.

V. Evaluation

Whereas the Assess Model task of the Modeling phase focuses on technical model assessment, the Evaluation phase looks more broadly at which model best meets the business and what to do next. This phase has three tasks:

Evaluate results:

Do the models meet the business success criteria? Which one(s) should we approve for the business? ### Review process: Review the work accomplished. Was anything overlooked? Were all steps properly executed? Summarize findings and correct anything if needed.

Determine next steps:

Based on the previous three tasks, determine whether to proceed to deployment, iterate further, or initiate new projects.

VI. Deployment

"Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise."

–CRISP-DM Guide

A model is not particularly useful unless the customer can access its results. The complexity of this phase varies widely. This final phase has four tasks:

Plan deployment:

Develop and document a plan for deploying the model.

Plan monitoring and maintenance:

Develop a thorough monitoring and maintenance plan to avoid issues during the operational phase (or post-project phase) of a model.

Produce final report:

The project team documents a summary of the project which might include a final presentation of data mining results.

Review project:

Conduct a project retrospective about what went well, what could have been better, and how to improve in the future.