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.

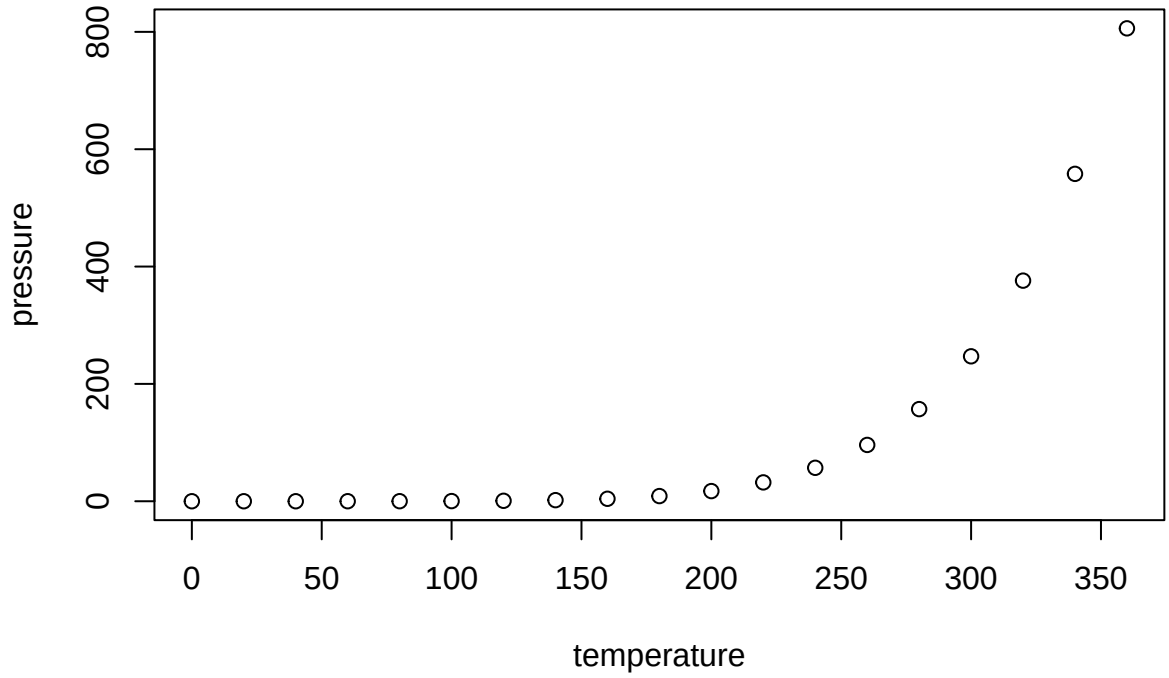
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
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

Newcastle University is a Russell group university offering high quality courses to students from around the world. Working with online partner Future Learn, the university are offering an online cyber security course. The course provider must show awareness of factors impacting on cohort recruitment such as demographics, age, and gender. The provider must review key performance indicators and analyse how appealing the course is to specific user groups. Course retention and understanding of students motive to enroll must be evaluated. In addition the provider must review barriers and challenges that sub groups of the cohort have experienced and how this may have prevented them completing the course. This will allow the provider to address these issues in future and develop a sustainable product. Academic outcomes of students are of interest to the course providers as this allows them to see what resources are having the greatest impact on learning but also allows them to identify areas where improved resources would impact on student outcomes. Finally the university must consider how it could keep up with changes in new technology and how students access the

content provided. Accessibility is a key factor and how the provider can ensure a course or product is more accessible to a wider audience, encouraging a greater enrolment globally.

The data files from 7 runs or repetitions of the course have been acquired. These files will be analysed to determine key links and trends within the data. The strength of the data collected will determine the strength of potential outcomes of this project. The data from the repetitions or reps needs to have similarities to ensure that files and scenarios are comparable. Clear identifiable relationships between the data files will help construct stronger lines of enquiry. Thsi will lead to outputs that will have a greater impact on informing future decision making within the process. From the third course rep information is included about video resources used within the course material and how and where the videos were accessed. Although this video data was not accessible for reps one and two this may be a way to compare if video files have had an impact since the introduction in rep three of the course. R scripts will be written within R studio software package to ensure that any work undertaken can be easily reproduced.

This project will look to meet the aims and goals set out in the business objectives. By exploring the data the intention of this project is to identify possible links between the introduction of video resources and the completion of the course. Is it possible to identify a groups of students who are more likely to complete the course and is there a clear reason for this? Are there clear indications why certain students fully complete the course or actually drop out of the course? By using the data to identify clear strengths and weaknesses of the course, this should inform decision making on what resources need strengthening or where to better support students in order to improve retention and academic outcomes.

R scripts will be used in order to clean and collate data. The use of R language allows the data not only cleaned but also means that the data can be displayed in clear visual formats that support what large datasets are saying. The R scripts will be accessed through R studio software that allows the scripts to be stored within a Project Template structure that means managing and accessing scripts is both organised and accessible. R studio then allows these scripts, visual data representations and charts to be merged together into a report that can be exported as a PDF file format that can be easily distributed. A key success criteria of this project is reproducibility and the CrispDM methodology gives a clear structure to the process and reproduction of the work taken place within the project. This methodology will be structured through the Project Template file structure to ensure that other users could replicate the work that had been undertaken.

A Gitlog will be included as part of the project to show version control techniques and ensure that changes to the work are recorded at regular intervals in the event of errors occurring. Version control allows all work to be rolled back to staged points in time.

II. Data Understanding

There are 53 .csv files provided by the university, spreading across seven repetitions of the course. The files all start with the prefix cyber.security then the number of the repetition of the course (cyber.security.1, cyber.security.2 etc). These are then split into eight possible datasets in the final five repetitions of the course and only seven possible datasets from the first two repetitions. This is because there is no "videostats" files until repetition 3 onwards. The files provided by the university are as follows:

Archetype survey responses
Enrolments
Leaving survey responses
Question responses
Step activity
Team members
Video stats
Weekly sentiment survey responses

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 modelling. 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.