CRISP-DM Project Writeup

Ben Jaeger

11 December, 2020

Contents

Chapter 1

Business Understanding

1.1 Determine Business Objectives

1.1.1 Background

The course was run seven times with start dates between 05/09/2016 and 10/09/2018 and a few datasets were made from each time a session was run. These data sets largely tracked the progress of the participants through the course, including their enrolment, question responses to quizzes, time taken per step of the session, etc. As more sessions were run, the content of the course and the data collected changed. The nature of the datasets provided is listed in the table below.

Dataset	Description	Sessions Used In
Archetypes	A data set containing the "archetypes" of a participant, 1 of 8 types	Session 1-7* **sessions 1, 2 are empty*
Enrollments	A data set containing lots of data on the participants including when they joined and left the course, gender, nationality, age range and others	Session 1-7
Leaving	A data set containing data about those who decided to leave the course early, when, why and how far they got	Session 1-7* **sessions 1, 2, 3 are empty*
Members	A data set containing data about those members running | the course (names have been expunged)	Session 2-7
Question Responses	A data set containing containing each participant's answers to the quiz questions as well as whether or not they're correct	Session 1-7
Step Activity	A data set containing data about each participant's progress through the steps of the course	Session 1-7
Sentiments	A data set containing the participants sentiments about the course week by week	Session 1-7* **sessions 1, 2, 3, 4 are empty*
Video Statistics	A data set containing lots of data about the videos participants watch during the course	Session 1-7* **sessions 1, 2 are empty*

1.1.2 Business Objectives and Success Criteria

This project was presented with no explicit questions to be answered or goals specified, thus the decision on what the ultimate goal of this project is will be reserved until some data has been explored and some insights gained.

1.2 Assess Situation

1.2.1 Inventory of Resources

Below are the resources available to the project:

Personnel:

- Ben Jaeger, Big Data CDT student at Newcastle University
- Joe Matthews, lecturer in statistics in the School of Mathematics, Statistics and Physics at Newcastle University*
- Matthew Forshaw, Senior Lecturer in Data Science and National Skills Lead to The Alan Turing Institute*

**Available to give advice/guidance only, no direct contributions will be made*

Data:

- All the data listed in section 1.1.1
- Any data or reference material found available freely online

Hardware:

- Ben Jaeger's personal PC

Software:

- R Studio (and any packages therein)
- Microsoft Excel
- Notepad++
- Mozilla Firefox

1.2.2 Requirements, Assumptions, and Constraints

Requirements
This report is to be read by a technically literate audience and so need not stray away from technical details. There are no legal constraints surrounding the data to be used. The results need not be important (indeed verifying a lack of surprising information is in a way a result) though they should be of a reasonable statistical rigour. The main aim of the project is more to develop and become comfortable with a suite of tools which allow us to extract interesting insights in a *quick*, *reliable* and *repeatable* manner. The project report is to be a maximum of 20 pages. A project presentation video of approx. 5 minutes also needs to be produced.

Assumptions
It is assumed the data set is complex but consistent enough to yeild interesting results. It is assumed the project can be completed on time*. It is assumed the intended audience of the report is of a high technical experience so little explanation of statistical concepts is needed. It is assumed all the requirements of the project (data-/software-/hardware-/knowledge-wise) are already in-hand or freely available.

**This was later proved false due to the mental health of the personnel invloved as mentioned in the Constraints section of section 1.2.2*

Constraints
The project is required to be finished by 23:45 on Friday 04/12/2020*. Several other projects of this kind are being run in parallel, while advice on small problems can be given no work may be shared "verbatim or in substance without specific acknowledgement" between them in accordance with Newcastle University's rules regarding plagiarism.

*This was later amended by an extension to 11/12/2020

1.2.3 Terminology

For consistencies sake the following terms are defined with the following meanings:

- Course - refers to the course taken as a whole without specifying a session
- Session - refers to a specific instance of the course being run (eg session 1 began 05/09/2016)
- Step - refers to a subsection of the course such as a quiz, article, video or other
- Participant - any person who took the course

1.3 Determine Data Mining Goals

1.3.1 Data Mining Goals and Success Criteria

As previously mentioned in section 1.1.2 this project was presented with no explicit objective or question to answer and so a decision on this is reserved until some data exploration has been done.

1.4 Produce Project Plan

1.4.1 Project Plan

The borad project plan is to do some initial data cleaning and exploration to gain a sense of what questions may be asked and answered using the dataset, a question or goal decided upon and then a few more cycles of cleaning and preparation until the data is fit to answer said question in whatever way seems fit, possibly graphically, possibly using some statisical tests, whatever is appropriate.

Chapter 2

Data Understanding and Preparation Cycle 1

2.1 Data Understanding

2.1.1 Collect Initial Data

Initial data collection conssited only of downloading a zip folder of CSVs and unzipping it, along with importing the data with project template

2.1.2 Describe Data

The datasets were given a quick descriptive overview in 1.1.1 but a more detailed look at the data in each is given in the table below:

(1) Archetypes

Columns	Description	Type
id	Numeric identifier, holds no specific info	Integer
learner_id	String identifier of a participant	String
responded_at	Datetime of info being received	Datetime
archetype	Archetype of participant, one of eight factors	String

(2) Enrolments

Columns	Description	Type
learner_id	String identifier of a participant	String
enrolled_at	Datetime of enrolment	Datetime
unenrolled_at	Datetime of unenrolment	Datetime
role	Participant role within course	String
fully_participated_at	Datetime of course completion	Datetime
purchased_statement_at	Datetime of statement purchase	Datetime
gender	Gender of participant	String
country	Reported country of participant	String
age_range	Age range of participant	String
highest_education_level	Education level of participant	String
employment_status	Employment status of participant	String
employment_area	employment area of participant	String
detected_country	Detected country of participant	String

(3) Leaving

Columns	Description	Type
id	Numeric identifier, holds no specific info	Integer
learner_id	String identifier of a participant	String
left_at	Datetime of participant departure	Datetime
leaving_reason	Given reason for departure	String
last_completed_step_at	Datetime of last completed step	Datetime
last_completed_step	Last completed step	String
last_completed_week_number	Last week completed	Integer
last_completed_step_number	Last step completed	Integer

(4) Members

Columns	Description	Type
id	Numeric identifier, holds no specific info	Integer
first_name	First name of member *(redacted for privacy)*	String
last_name	Last name of member *(redacted for privacy)*	String
team_role	Team role of member	String
user_role	User role of member	String

(5) Quiz Responses

Columns	Description	Type
learner_id	String identifier of a participant	String
quiz_question	Question asked	String
question_type	Type of question	String
week_number	Week in which question was asked	Integer
step_number	Step in which question was asked	Integer
question_number	Number of question	Integer
response	Response given by participant	String
cloze_response	Response given by participant	String
submitted_at	Datetime of question being answered	Datetime
correct	Correct/incorrect answer?	Boolean

(6) Step Activity

Columns	Description	Type
learner_id	String identifier of a participant	String
step	Step being accessed	String
week_number	Week of step being accessed	Integer
step_number	Number of step being accessed	Integer
first_visited_at	Datetime of first access	Datetime
last_completed_at	Datetime of completion	Datetime

(7) Sentiments

Columns	Description	Type
id	Numeric identifier, holds no specific info	Integer
responded_at	Datetime of sentiment submission	Datetime
week_number	Week being reviewed	Integer
experience_rating	Rating given for week of course	Integer
reason	Reason given for rating	String

(8) Video Stats

Columns	Description	Type
step_position	Step of course for the video	String
title	Title of video	String
video_duration	Length of video	Integer
total_views	Total views	Integer
total_downloads	Total downloads	Integer
total_caption_views	Total captions viewed	Integer
total_transcript_views	Total transcript viewed	Integer
viewed_hd	Times viewed in HD	Integer
viewed_five_percent	Percentage viewers who viewed 05% of video	Decimal
viewed_ten_percent	Percentage viewers who viewed 10% of video	Decimal
viewed_twentyfive_percent	Percentage viewers who viewed 25% of video	Decimal
viewed_fifty_percent	Percentage viewers who viewed 50% of video	Decimal
viewed_seventyfive_percent	Percentage viewers who viewed 75% of video	Decimal
viewed_ninetyfive_percent	Percentage viewers who viewed 95% of video	Decimal
viewed_onehundred_percent	Percentage viewers who viewed 100% of video	Decimal
console_device_percentage	Percentage viewers who viewed on console	Decimal
desktop_device_percentage	Percentage viewers who viewed on desktop	Decimal
mobile_device_percentage	Percentage viewers who viewed on mobile	Decimal
tv_device_percentage	Percentage viewers who viewed on tv devices	Decimal
tablet_device_percentage	Percentage viewers who viewed on tablet	Decimal
unknown_device_percentage	Percentage viewers who viewed on unknown device	Decimal
europe_views_percentage	Percentage viewers in Europe	Decimal
oceania_views_percentage	Percentage viewers in Oceania	Decimal
asia_views_percentage	Percentage viewers in Asia	Decimal
north_america_views_percentage	Percentage viewers in N. America	Decimal
south_america_views_percentage	Percentage viewers in S. America	Decimal
africa_views_percentage	Percentage viewers in Africa	Decimal
antarctica_views_percentage	Percentage viewers in Antarctica	Decimal

2.2 Data Preparation

2.2.1 Select Data

Due to this being the inital cycle of data cleaning and preparation, no data has been excluded at this stage

2.2.2 Rationale for Inclusion/Exclusion

There is no reason to exclude data which may lead to insights, that said, as mentioned in 1.1.1 some data sets have literally no data in them, thus these data sets have been ignored and won't be imported for cleaning and analysis as they just clutter the workspace. Additionally, the sentiments dataset for session 5 has only one row in it and has thus also been excluded along the same lines.

2.2.3 Clean Data

Due to this being the first round of data cleaning there were a fair few cleaning steps taken.

The first, relatively minor, step taken was to rename the data imported from the automatic names given by ProjectTemplate to more readable names, they were renamed in the manner laid out in the table below: *<num> is a stand-in for the session number in the dataset name*

Renamed from	Renamed to
• cyber.security.<num>_archetype.survey.responses	• archetype_<num>
• cyber.security.<num>_enrolments	• enrolments_<num>
• cyber.security.<num>_leaving.survey.responses	• leaving_<num>
• cyber.security.<num>_question.response	• qresponses_<num>
• cyber.security.<num>_step.activity	• sactivity_<num>
• cyber.security.<num>_team.members	• members_<num>
• cyber.security.<num>_video.stats	• vidstats_<num>
• cyber.security.<num>_weekly.sentiment.survey.responses	• sentiments_<num>

Next up was cleaning the data types of the columns in the data frames. The first of these was to entries into proper NA values rather than an empty string or "unknown", etc.

The second area of cleaning was the datetime columns that appear in many of the datasets, these were all in string form which makes mathematical procedures on such data impossible, thus they were turned in a POSIXct type, which stores the datetime as seconds from 00:00 01/01/1970.

```
str_datetime = "2017-10-04 09:23:14 UTC"
typeof(str_datetime)
```

```
## [1] "character"
```

```
posixct_datetime = as.POSIXct(str_datetime, format = "%F %T UTC", tz = "GMT")
typeof(posixct_datetime)
```

```
## [1] "double"
```

Unfortunately, `as.POSIXct()` is not a vectorised function and so the conversions must be done one by one, additionally there are a lot of such conversions that need doing if it is to be done to all the datasets so this munge file was commented out after being made until the datasets being considered were reduced.

The next item for cleaning were any columns with category factors, eg, the reason for leaving is one of 8 possible options that are consistent across all the sessions. (The opportunity was also taken to correct an encoding error where an apostrophe was encoded as "â€™").

2.3 Data Understanding and Preparation Cycle Summary

2.3.1 Summary of Work Done

A lot of work was done in this cycle both with regards to exploring the data and cleaning it for future analysis. After doing all this, it was decided that the `learner-id` column was pivotal with regards to linking between datasets.

2.3.2 Future Plans

After the exploration done it was decided that the most interesting avenue would be to explore how a factor like gender affected people scores on the quiz questions, the main datasets involved in this investigation would be the `enrolments` datasets as they hold information about the participants, the other would be the `quiz responses` datasets as they hold information on the participant performance. The main challenge in combining these data sets would be to see how effectively the `learner-id` columns could be mapped onto each other.

Chapter 3

Data Understanding and Preparation Cycle 2

3.1 Data Understanding

3.1.1 Explore Data

The first goal of this round of exploration and cleaning was to see if the `learner-id` columns of the `enrolments` and `quiz responses` datasets could be mapped onto each other, to explore this the `merge` function in R was looked into, this function performs merges on dataframes much like a merge in SQL. `Merge` is able to perform inner, left/right outer and full outer joins, for this operation an inner join was decided to be the required version as having a participant entry without a quiz response or vice versa was unneeded and would just clutter the dataset. It was also decided that much of the two datasets could be left out in the merge, details of which are discussed in sections 3.2.2 and 3.2.3.

3.2 Data Preparation

3.2.1 Select Data

Firstly, since it was decided that only the `enrolments` and `quiz responses` datasets were needed, all other data sets could be ignored from being imported when loading the project, this was achieved by adding the line `data_ignore: /archetype/, /leaving/, /weekly-sentiment/, /members/, /step-activity/, /video/` to the config, this uses regex to match any string between two forward slashes to select which datasets to ignore.

3.2.2 Integrate Data

The first job in merging the datasets was to retrieve the data and define the parameters of the merge, by default `merge` performs inner merges. To automate the process of merging the datasets a variable was used to iterate over 1 to 7 (for the seven sessions run), then the `get` function was used to retrieve the data. `Get` takes a string name of any environment object and returns that object (interestingly, this is a "pass by value" operation so the original is preserved). Lastly the new dataset needed to be assigned to a new variable, for this the `assign` function was used, this takes a name and a value, and creates a new variable with that name and value. Combining the `get`, `merge` and `assign` functions in a for loop the below code was made:

```
for (i in 1:7) {
  enrol_df = get(paste0("enrolments_",i))
  qresp_df = get(paste0("qresponses_",i))

  new_df = merge(x = enrol_df, y = qresp_df, by.x = "learner_id", by.y = "learner_id")
```

```
  assign(paste0("enANDqr_",i), new_df)
}
```

Which takes the `enrolments` and `question responses` datasets for a given session, merges them on the `learner_id` column with an inner join and stores the result in a new dataframe.

3.2.3 Format Data

Given that analysis was being done on how gender might affect quiz performance for participants, many of the columns in the new dataset could be trimmed out. In addition a new column was added to track which session the question was asked in, this resulted in the edited code chunk below:

```
for (i in 1:7) {
  enrol_df = get(paste0("enrolments_",i))
  qresp_df = get(paste0("qresponses_",i))

  new_df = merge(x = enrol_df, y = qresp_df, by.x = "learner_id", by.y = "learner_id")
  new_df_trim = new_df[colnames(new_df)[c(1,7,14,22)]]
  new_df_trim = cbind(new_df_trim, data.frame(session = i))

  assign(paste0("enANDqr_",i), new_df_trim)
}
```

The columns preserved in this trimming are discussed in section 3.2.4

Additionally a dataset containing all the sessions in one was made.

3.2.4 Dataset Description

During the process of making the merged datasets, many columns were left out and only a few were included, below is the full list of columns from both and whether they were preserved:

Column	Dataset of Origin	Included?
learner_id	Both	Yes
enrolled_at	Enrolments	No
unenrolled_at	Enrolments	No
role	Enrolments	No
fully_participated_at	Enrolments	No
purchased_statement_at	Enrolments	No
gender	Enrolments	Yes
country	Enrolments	No
age_range	Enrolments	No
highest_education_level	Enrolments	No
employment_status	Enrolments	No
employment_area	Enrolments	No
detected_country	Enrolments	No
quiz_question	Quiz Responses	Yes
question_type	Quiz Responses	No
week_number	Quiz Responses	No
step_number	Quiz Responses	No
question_number	Quiz Responses	No
response	Quiz Responses	No
cloze_response	Quiz Responses	No
submitted_at	Quiz Responses	No
correct	Quiz Responses	Yes
session	Neither	Yes

3.3 Data Understanding and Preparation Cycle Summary

3.3.1 Summary of Work Done

By taking advantage of `merge` new data sets were made for each session combining the `enrolments` and `quiz responses` datasets, the new data sets holding the columns described in 3.2.3. A dataset combining all of the sessions together was made too.

3.3.2 Future Plans

With the data sets made the plan for the analysis of how gender affects quiz performance is to do a visual inspection, comparing the proportion of correct/incorrect results as well as performing t-tests to see if the mean percentage correctness varies in a statistically significant way.

Chapter 4

Modeling

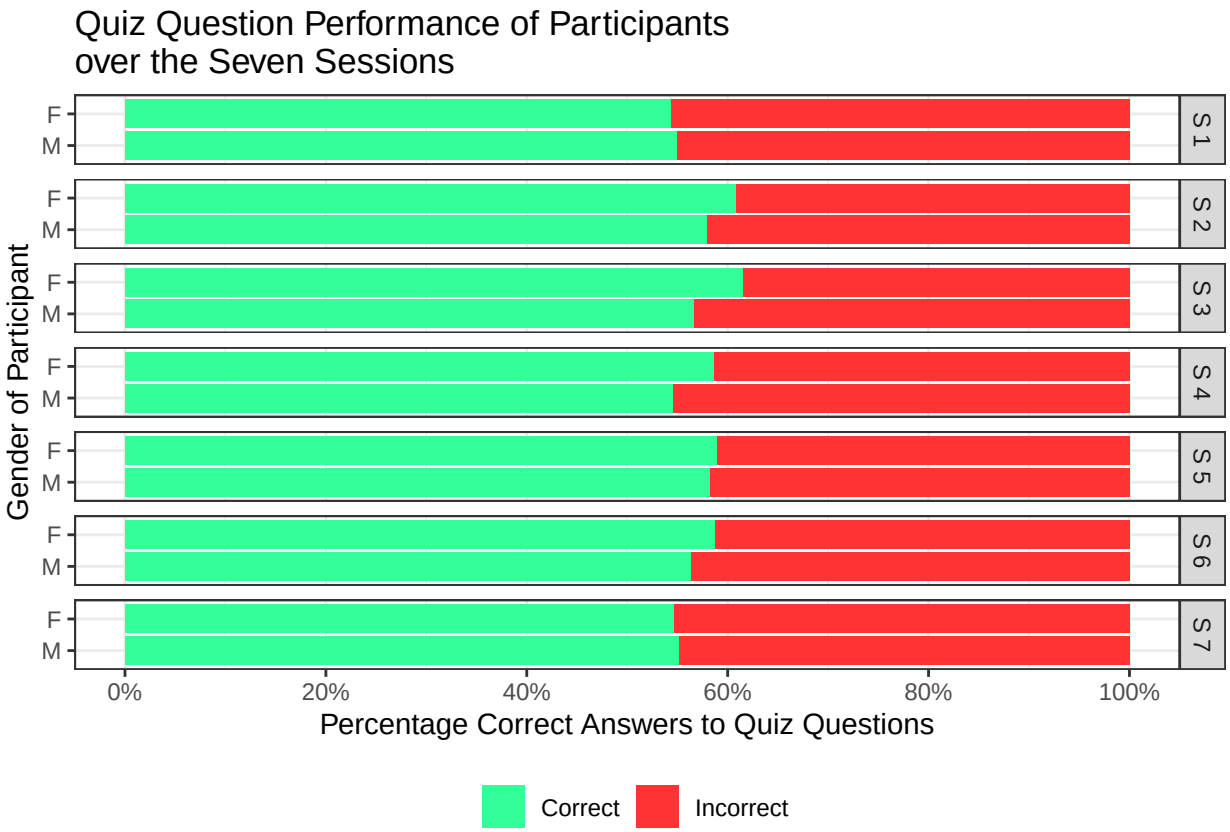
4.1 Visual Inspection of Data

4.1.1 Graphed Data

The first sensible step before investing time a statistical investigation was to perform a graphical exploration of the data, to this end the proportion of correct to incorrect answers for male and female participants were compared side by side over the seven sessions. The below code was used

```
enANDqr_all %>%
  filter(!is.na(gender)) %>%  # filter NAs
  filter(gender == "male" | gender == "female") %>%  #filter non binary and other
  ggplot() +
    geom_bar(aes(x = factor(gender), fill = correct), position = "fill") +  # filled barchart
    # faceted on sessions
    facet_grid(rows = vars(session), labeller = labeller(session = labels)) +
    labs(x = "Gender of Participant", y = "Percentage Correct Answers to Quiz Questions",
         title = "Quiz Question Performance of Participants \nover the Seven Sessions",
         fill = "") +  # nice labels, note: blank legend title
    scale_x_discrete(labels = c("M","F")) +  # shorter than male/female
    # percentage labels look nicer
    scale_y_continuous(breaks = seq(0,1,by=0.2),labels = paste0(seq(0,100,by=20), "%") ) +
    #correct = green, incorrect = red
    scale_fill_manual(values = c("#FF3333", "#33FF99"), labels = c("Incorrect", "Correct")) +
    coord_flip() +  #horizontal layout looks a little nicer
    guides(fill = guide_legend(reverse = TRUE)) +  # show correct on left
    theme_bw() +  # best theme
    theme(legend.position = "bottom")  # legend at bottom
```

Producing the following plot:



4.1.2 Graphical Implications

From this graph it can be seen there is some distinction between the performance of men and women in quizzes across the seven sessions, with this we have some justification to perform a statistical test on the data.

4.2 Select Statistical Techniques

4.2.1 Statistical Technique

The chosen statistical test was a t test, used to see if the means between two groups varies in a significant way. To make use of this test, for every question in a set, the percentage correct answers for men and women were found, then a paired t test can be performed taking the pairs as those percentages for men and women respectively. The code for this is shown below:

```
for (i in 1:7){
  m_percents = c()
  f_percents = c()

  df = get(paste0("enANDqr_", i))

  for (q in unique(df$quiz_question)) {
    m_total = length(df$correct[na.omit(df$quiz_question == q & df$gender == "male")])
    m_corr = sum(df$correct[na.omit(df$quiz_question == q & df$gender == "male")] == "true")
    m_percents = c(m_percents, m_corr/m_total)

    f_total = length(df$correct[na.omit(df$quiz_question == q & df$gender == "female")])
    f_corr = sum(df$correct[na.omit(df$quiz_question == q & df$gender == "female")] == "true")
    f_percents = c(f_percents, f_corr/f_total)
  }
```

```
  print(t.test(x=m_percents, y=f_percents))

}
```

4.2.2 Assumptions Made

In the t-test comparing the means of two independent samples, the following assumptions should be met:

1. The means of the two populations being compared should follow Normal distributions.

This is assumed as true since each percentage is based on the total correct answers which can be though of as a sum of Bernoulli trials for each time the question is answered, under the Central Limit Theorem, this averaging results in a Normal variable.

2. If using Student's original definition of the t-test, the two populations being compared should have the same variance.

Welch's t-test is insensitive to equality of the variances regardless of whether the sample sizes are similar, thus this is not a concern.

3. The data used to carry out the test should either be sampled independently from the two populations being compared or be fully paired.

This is assumed as true since any participants answer is independent from any others, thus their distributions are too.

4.3 Implement Technique

4.3.1 Technique Results

Running the above code we find the below results:

Session	Male Mean	Female Mean	T Statistic	P Value	Significant at a 10%/5%/1% level?
1	0.549	0.550	-0.157	0.876	No/No/No
2	0.593	0.689	-1.573	0.125	No/No/No
3	0.563	0.557	0.184	0.855	No/No/No
4	0.552	0.545	0.192	0.850	No/No/No
5	0.598	0.593	0.102	0.920	No/No/No
6	0.498	0.599	-1.507	0.141	No/No/No
7	0.582	0.549	1.053	0.299	No/No/No

4.4 Assess Technique Results

4.4.1 Results Assessment

As we can plainly see in the above table, there is no significant evidence to believe men or women do better on the quizzes at any reasonable level.

Chapter 5

Evaluation

5.1 Evaluate Results

5.1.1 Assessment of Data Mining Results w.r.t. Business Success Criteria

Given that the project was presented without a specific goal, that criterion isn't available for comparison. Against the goal that was set in section 2.3.2, evaluating how gender affects quiz performance, it is very fair to say the project has succeeded in finding a statistically rigorous result.

5.2 Review Process

5.2.1 Review of Process

The overall process of finding this result was reasonably efficient given it was a goal that wasn't set in the beginning, some work could have been streamlined, the munging of datetimes for example was ultimately unnecessary. However, it can be argued, given the goal which ultimately didn't need that work hadn't been set, the work done might have been worthy in some other world where some other goal was set.

5.3 Determine Next Steps

5.3.1 Next Steps

With these results found, the next steps are to:

- Finish this report
- Write a report on how CRISP-DM was implemented in this project and how it went
- Record an oral presentatiopn of the work done
- Collect the relevent files together for submission

Chapter 6

Deployment

6.1 Plan Monitoring and Maintenance

6.1.1 Monitoring and Maintenance Plan

There is no need to give this work any ongoing attention.

6.2 Produce Final Report

6.2.1 Final Report

This is the final report.

6.2.2 Final Presentation

A video presenting the results will be made. The video is to be approximately 5 minutes long and will look at the work done from the perspective of showing the results to the stakeholders.

6.3 Review Project

6.3.1 Experience Documentation

A short report with my thoughts on CRISP-DM will also be produced.