# CRISP-DM Project Writeup

Ben Jaeger

10 December, 2020

# Contents

# Chapter 1

# Overview

# Chapter 2

# Business Understanding

## 2.1 Determine Business Objectives

### 2.1.1 Background

The data provided comes from a course run by FutureLearn about computer security. The participants on the course come from a wide range of backgrounds, some are university students, some are from the private or public sectors, there is a wide range of age brackets that they can belong to and they can come from a variety of countries, among other factors. 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*<br>**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*<br>**sessions 1, 2, 3 are empty* |
| Members | A data set containing data about those members of FutureLearn 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* |

| Dataset | Description | Sessions Used In |
|---|---|---|
| Video Statistics | A data set containing lots of data about the videos participants watch during the course | Session 1-7*<br>**sessions 1, 2 are empty* |

### 2.1.2 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.

## 2.2 Assess Situation

### 2.2.1 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 2.1.1
- 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

### 2.2.2 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 2.2.2*

**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*

### 2.2.3 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

## 2.3 Determine Data Mining Goals

### 2.3.1 Data Mining Goals and Success Criteria

As previously mentioned in section 2.1.2 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.

## 2.4 Produce Project Plan

### 2.4.1 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 3

# Data Understanding and Preparation Cycle 1

## 3.1 Data Understanding

### 3.1.1 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

### 3.1.2 Describe Data

The datasets were given a quick descriptive overview in 2.1.1 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 |

## 3.2 Data Preparation

### 3.2.1 Select Data

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

### 3.2.2 Rationale for Inclusion/Exclusion

There is no reason to exclude data which may lead to insights, that said, as mentioned in 2.1.1 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.

### 3.2.3 Clean Data

Due to this being the first round of data cleaning there were many 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€™”)

## 3.3 Data Understanding and Preparation Cycle Summary

### 3.3.1 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.

### 3.3.2 Future Plans

After the exploration done it was decided that the most interesting avenue would be to explore how various factors affected people scores on trhe quiz questions, the main datasets involved in this investigation would be the `enrolments` datasets as they hold information about the particpants, 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 4

# Data Understanding and Preparation Cycle 2

## 4.1 Data Understanding

### 4.1.1 Collect Initial Data

#### 4.1.1.1 Initial Data Collection Report

### 4.1.2 Describe Data

#### 4.1.2.1 Data Description Report

### 4.1.3 Explore Data

#### 4.1.3.1 Data Exploration Report

### 4.1.4 Verify Data Quality

#### 4.1.4.1 Data Quality Report

## 4.2 Data Preparation

### 4.2.1 Select Data

#### 4.2.1.1 Rationale for Inclusion/Exclusion

### 4.2.2 Clean Data

#### 4.2.2.1 Data Cleaning Report

### 4.2.3 Construct Data

#### 4.2.3.1 Derived Attributes

#### 4.2.3.2 Generated Records

### 4.2.4 Integrate Data

#### 4.2.4.1 Merged Data

### 4.2.5 Format Data

#### 4.2.5.1 Reformatted Data

### 4.2.6 Dataset

#### 4.2.6.1 Dataset Description