CSC8631 Assignment Report

Marc Birkett

07/11/2021

CSC 8631 - Data Management and Exploratory Data Analysis

An investigation into an MDOC dataset using a CRISP-DM model. The model is the industry standard approach to data mining projects, it has six phases and is iterative. It is described as a "set of guardrails to help you plan organise and implement your data science (or machine learning) project."(Data Science Alliance, 2021).

The phases and tasks of a CRISP-DM project are outlined by the Data Science Alliance as per:

1. Business Understanding - what does the business need?

- Understand the business objective. What does the customer want to accomplish?
- Assess the situation. Determine resource availability, requirements, risks and contingencies
- Data Mining Goals - what does success look like?

2. Data Understanding - what data do we have / need? Identify, collect and analyse.

- Collect the initial data, aquire and load into data analysis tool.
- Describe the data, examine and document properties such as data format, number of records and identities
- Explore the data, query and visualise and idetntify the relationships.
- Verify the data quality and document any issues.

3. Data Preparation - prepare the final datasets for modelling.

- Determine which data set to be used and reasons for inclusion / exlusion.
- Clean the data
- Construct data, carry out any required feature engineering.
- Integrate data, create new datasets by combining data sources
- Re-format data as necessary, cast data types as required.

4. Modelling - what modelling techniques should we apply?

- Determine which models to try.
- Design tests.
- Build and assess the model.

5. Evaluation - which model best meets the business objectives?

- Evaluate the models against the business success criteria identified in step 1.

- Review the processes, was anything overlooked and everything properly executed?
- Determine next steps, what do we need to do to deploy.

6. Deployment - how do stakeholders access the results?

- Plan deployment, document.
- Plan monitoring and maintenance.
- Produce the final report.

This report will cover the first three steps of this model and a minimum of two iterations. I will attempt to implement all of the associated tasks. With the lack of formal requirements provided by the assignment it will be difficult to define a business need in the first iteration of the model therefore tasks such as defining the business objective and data mining goals will be defined for the second iteration as a result of the first. So, each therefore the two iterations of the model will be:

1. Iteration 1

- Data Understanding
- Data Preparation

2. Iteration 2

- Business Understanding - define and understand the hypothesis from Iteration 1.
- Data Understanding
- Data Preparation.

The investigative project will be completed in R using R Studio and has been set up using ProjectTemplate to provide structure and repeatability. Version control is provided by Git and this report created with R Markdown. Various libraries have been imported and used from the Tidyverse regarding data import and management (Dplyr, Readr) and visualisation (GGPlot2).

Iteration 1

A first iteration completing the understanding and preparation step will be carried out. The Future Learn MDOC dataset was downloaded as a zip file and reviewed. The data was supplied in csv files covering 8 different areas of the software over multiple stages, this was imported into R into 8 data frames from the original for easier analysis. The data was then reviewed.

Step 2: Data Understanding Upon initial import is was found that the detected data types were not consistent and so his was dynamically set on import. Particularly any ID property was set to be an integer and date time properties we converted to the local POSIX time format.

Upon investigation it was decided that some data would be disregarded due to lack of breath. Therefore, we have data covering the life cycle of a student including:

- **Enrollments**. Enrollment is a categorical dataset with n=37296 items and p=14 variables. It was found that the majority of data for all properties was "unknown" and therefore not suitable for further analysis.

  - **Candidate Keys** - "Learner_ID" of character data type.

- **Fields and data types** - Properties include Country, Gender, Age Range, Highest Education Achieved, Employment Area and Employment Status, Role all of type character. Also date / time of enrolled, un-enrolled, fully participated and purchased statements.
- **Related data sets** - Team Member, Leaving Survey and Step Activity can be related on Learner_ID
- **Data Quality** - The data types across date times in multiple files were inconsistent and therefore they were cast on import. The vast majority of data for each categorical value is "Unknown". Once the data was imported a basic type and missingness check was carried out. This showed type failures of 1 record on each of the date fields (enrolled, unenrolled, fully participated and purchased statement), there were also significant gaps in the same date based data, between 35k and 37k records. However, there were no issues with the categorical data. This was then plotted to investigate the data as shown in Figure 1 "Categorical Enrollment Data".
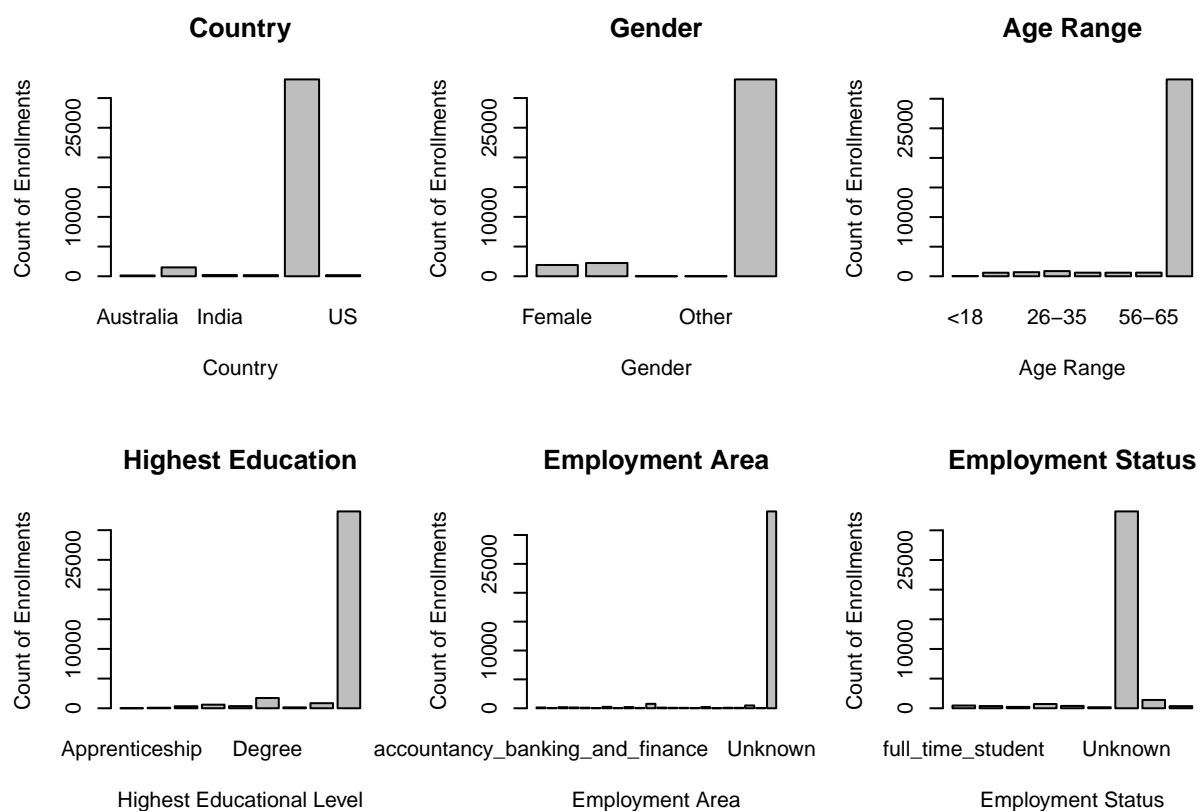

Figure 1: Categorical Enrollment Data

It can be seen that the majority of the data for the categorical data fields is "unknown" and therefore not suitable for further analysis.

- **Step Activity**. This indicates the stage in the course that the related data occurred at. It is principally categorical data with n=423072 items and p=6 variables.

  - **Candidate Keys** - A compound key of "Learner_ID" of character data type and step (integer). This identifies data about the step the student was at.
  - **Fields and data types** - "Learner_ID" of type character, "week_number" and "step_number" both of which are integer and "step" which is also an integer and is a concatenation of the

two. "First_Visited_At" and "Last_Completed_At" are both date times and indicate when the student was at that stage.

– **Related data sets** - Enrollments via "Learner ID", Question Response, Video Stats and Sentiment Survey via Step. Each step in the step activity data frame has associated quiz questions and responses.

– **Data Quality** - Data types were consistent across the files to be imported. Date times were cast on import. Basic data quality checks indicate that "First_Visited_At" and "Last_Completed_At" are natively characters and so they were cast to date times. Significant gaps were observed on "last_completed_at" which may be correct.

- **Leaving Survey**. The leaving survey responses comprises of categorical, character and date based data regarding feedback from individuals which have left the course. The data frame has n=403 items and p=10 variables. The primary key of this table is ID which was cast to an integer in the import process.

  – **Candidate Keys** - "ID" which was cast to integer during import.

  – **Fields and data types** - "ID" as integer, "Learner_ID" as character, "Left_at" as datetime, the date and time a student left the course, "leaving_reason" as characters, "last_completed_step" as character and "last_completed_step" as as date time, also the "last_completed_week_number" and "last_completed_step_number".

  – **Related data sets** - Enrollments on "Learner_ID". Investigation into the key fields of "last_completed_step", "last_completed_step_number" and "last_completed_week_number" revealed that "last_completed_step" was a concatenation of the other two fields and could be used to relate to the Step Activity data. This allows the identification of the stage that students leave the course. This can be seen below.

```
## # A tibble: 5 x 3
##   last_completed_step last_completed_step_number last_completed_week_number
##   <chr>                                    <dbl>                      <dbl>
## 1 1.3                                          3                          1
## 2 1.19                                        19                          1
## 3 3.18                                        18                          3
## 4 1.3                                          3                          1
## 5 1.2                                          2                          1
```

  – **Data Quality** - Upon import it was necessary to standardise the data type across multiple files, id was cast to integer, last completed step to character and the last completed step and week numbers to integer. All date times were also cast to date time. Upon investigation of the data it was found that there were multiple leaving reasons which amounted to "lack of time" (Figure 1), this field was merged to make it comparable to other values (Figure 2)

```
##
##                      I don't have enough time
##                                            41
##                          I prefer not to say
##                                            18
##                                         Other
##                                            48
## The course required more time than I realised
##                                            16
##                      The course was too easy
##                                            10
##                      The course was too hard
```

4

```
##                                                      12
##            The course wasn't what I expected
##                                                      20
##        The course won't help me reach my goals
##                                                      21
```
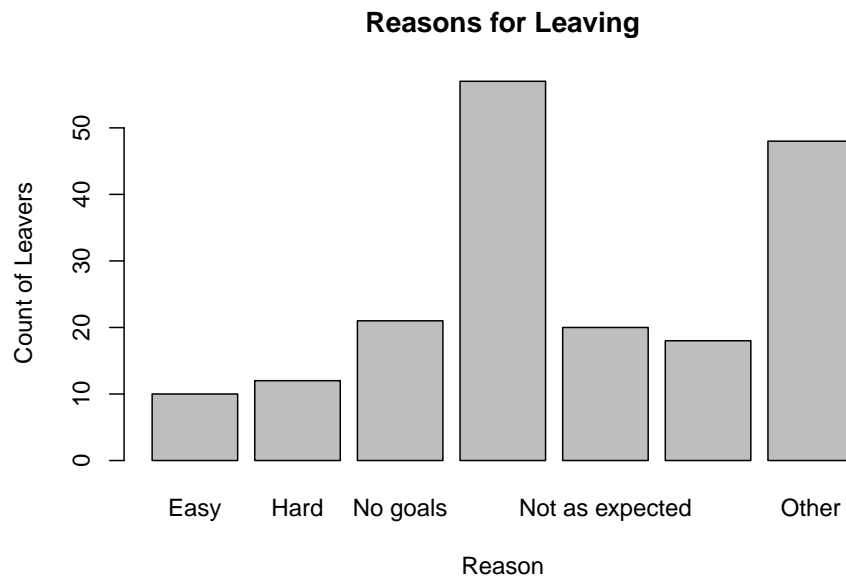
Once corrected the data is comparable as per below.

Reasons for Leaving

Figure 2: Engineered Reasons for Leaving

- **Question Responses**. Each step in the step activity data frame has associated quiz questions and responses. This is stored in the questions data frame. The question responses data is categorical data with n=176463 items and p=10 variables.

    - **Candidate Keys** - "Learner_ID" of character data type and quiz_question which is a concatenation of week_number, step_number and question_number. By concatenating just the week_number and step_number we are able to calculate the question step and relate to the Step Activity data set.

    - **Fields and data types** - Other properties outside of the keys are "question_type" which is exclusively "MultipleChoice", response of character which is the answer to the question, cloze_response which is always NA, submitted_at which is a datetime of the submitted at date and correct which is a boolean. This would give us access to the scores of each student

    - **Related data sets** - Some feature engineering would allow us to relate to the Step Activity data set and the learner_id allows us to relate to the Enrollments data set and any associated Leaving Surveys.

    - **Data Quality** - Simple data quality checks were carried out which showed the correct data types, and that Cloze_Response is always null. A review of week_number, step_number and question_number confirmed that they could be concatenated to form the tables primary key and a step key, as per Figure 4 below, "Question Response Key Construct".

```
## # A tibble: 29 x 4
```

```
## # Groups:   quiz_question [29]
##    quiz_question week_number step_number question_number
##    <chr>               <dbl>       <dbl>           <dbl>
##  1 1.7.1                   1           7               1
##  2 1.7.2                   1           7               2
##  3 1.7.3                   1           7               3
##  4 1.7.4                   1           7               4
##  5 1.7.5                   1           7               5
##  6 1.7.6                   1           7               6
##  7 3.11.1                  3          11               1
##  8 3.11.2                  3          11               2
##  9 3.11.3                  3          11               3
## 10 2.8.1                   2           8               1
## # ... with 19 more rows
```

It can be clearly seen that concatenating the columns gives us the key.

- **Video Views**. Each step in the step activity data frame also has associated video view data. The video stats data is multivariate data with n=65 items and p=29 variables. There are significantly more data variables in this relation then in those examined previously, and significantly less data items.

    - **Candidate Keys** - "Step_Position" identifies the videos position within in the course as a whole and could be related to Step Activity. This along with the video title would uniquely identify he video and the rest of the data in the data set is already grouped.
    - **Fields and type** - All of the fields in this dataset other than "title" are categorical and are integers. The data in this relation can be subsetted into data pertaining to percentage complete, type of device and location of the video view which allows investigation of each specific facet. This will be revisited at the data preparation stage if required.
    - **Related data sets** -The step_position allows us to relate to step_activity which allows relationships through the rest of the dataset.
    - **Data Quality** - Data quality checks for type and missingness did not raise any issues. Upon investifation, the table is wide, it will be necessary to pivot to see any meaningful data. Also the data is expressed as percentages, to make the data comparable it will be necessary to convert to absolute numbers, and compare back to total_views.

- **Weekly Sentiments**. This is completed by each student per week. The sentiment data has n=181 items and p=6 variables. This table has a categorical properly rating and character data to cover the sentiment. I will not be looking to cover sentiment analysis in this work so only the categorical value will be looked at.

    - **Candidate Keys** - "ID" has been cast to integer during the import process, other than week_number i do not see a key to create a relationship into the wider data set.
    - **Fields and data types** - "Experience_rating" is an integer between 1 and 3 which gives the rating of the student for that week. "Week_number" is the number for the week between 1 and 3 and "reason" is a character string to justify the experience_rating.
    - **Related data sets** - None
    - **Data Quality** - ID, Week number and experience rating were cast to integer on import to standardise the data type across multiple data sets. Responded at was also cast from character to date time. Data quality checks revealed correct data types, and reason being missing in 111 cases from 181. There wwold be value in coverting this data to percentages.

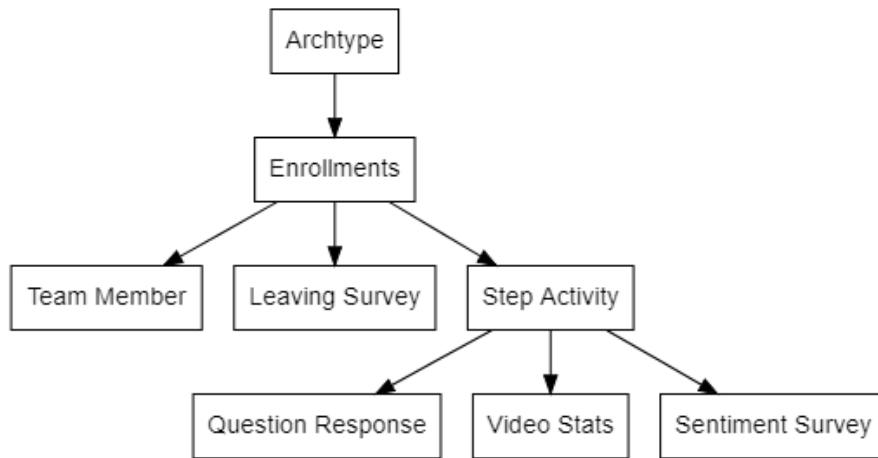The relationships between data sets as discussed above are documented below.

Figure 3: Relationships between data frames.

Data Preparation   Each set of data was imported into a dataframe during the data understanding stage. This was done by filtering the file set and reading the contents of each file to a list. This was then bound by Dplyr. The example below binds all *question.response.csv files into a dataframe.

```r
#working files
files = list.files(pattern="*question.response.csv")

datalist = lapply(files, function(i){
  csv <- read_csv(i, show_col_types = FALSE)
  csv$stage_id <- substring(i,16,16)
  csv
})
dfQR <-dplyr::bind_rows(datalist)
```

As we still do not have a clear business objective in this first iteration, the data preparation stage will be used to implement any data changes to make a more completed and integrated dataset. The changes implemented here may or may not be required in the analysis carried out in iteration 2, they have merely been observed while carrying out the steps during the data understanding process. Processes such as re-formatting and casting the data have already been carried out in the data understanding step to successfully merge multiple CSV files into one data frame.

Particularly, it was identified that some feature engineering could be carried out to prepare the data for further analysis. These fixes were implemented in the munge process. Particularly:

- **Step Activity**, this dataset has the step start and end date, some data engineers was completed to add an "isComplete" flag to the dataset, as well as a "completedTime" integer indicating the difference between the start and end date of the step.

    - An isComplete flag was added by checking the last_completed_at date for NA.

```r
dfSA$isComplete = !is.na(dfSA$last_completed_at)
```

- A time to complete count in days was added by subtracting the start date from the end date.

```r
dfSA$timeToComplete = difftime(dfSA$last_completed_at, dfSA$first_visited_at,
                               units="days") #calculate the difference
```

- Question Responses, it was identified that while the dataset is missing the step column to relate the Step Activity table, this can be created from data which is present. This allows us to integrate responses to steps and to the wider dataset.

  - The quiz_question key was split on . to a list. SApply was then used to concatenate the first two elements together split by a dot. This was then added to the dataset.

- Video Views, it was identified that this dataset has a logical collection of three types of data and could be split into three datasets, it was all observed that a pivotted data set would be required for analysis and the conversion of data from "percentages of percentages" to absolute counts.

  - Splitting out the child dataset was carried out by taking a slice of the original dataset.

  ```r
  dfVSTotals = dfVS[,c(1,2,4, 9:15)]
  ```

  - The data frame was then extended by adding the appropriate properties. In this example a helper function was used to express the percent as raw numbers..

  ```r
  dfVSTotals$"05" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_five_percent)
  dfVSTotals$"10" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_ten_percent)
  dfVSTotals$"25" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_twentyfive_percent)
  dfVSTotals$"50" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_fifty_percent)
  dfVSTotals$"75" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_seventyfive_percent)
  dfVSTotals$"95" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_ninetyfive_percent)
  dfVSTotals$"99" = as.percent(dfVSTotals$total_views, dfVSTotals$viewed_onehundred_percent)
  ```

  - The result was then selected using Dplry to produce a neater table. A pivotted table was then created to support reporting. The figure below shows the transformed data in a logical structure for reporting. This was carried out for all three datasets.

  ```r
  dfVSTotals = select(dfVSTotals, step_position, title, "05", "10", "25", "50", "75", "95", "99")
  dfVSTotalsPivot = dfVSTotals %>%
    pivot_longer(!(1:2), names_to = "percentviewed", values_to = "count") #createpivot
  ```

The formatted data for reporting is as per below.

```
## # A tibble: 455 x 4
##    step_position title                         percentviewed count
##            <dbl> <chr>                         <chr>         <int>
## 1            1.1  Welcome to the course        05             1276
## 2            1.1  Welcome to the course        10             1250
## 3            1.1  Welcome to the course        25             1218
## 4            1.1  Welcome to the course        50             1167
## 5            1.1  Welcome to the course        75             1130
## 6            1.1  Welcome to the course        95             1102
## 7            1.1  Welcome to the course        99             1056
## 8            1.14 Why would anyone want your data? 05          660
```

```
## 9           1.14 Why would anyone want your data? 10              645
## 10          1.14 Why would anyone want your data? 25              623
## # ... with 445 more rows
```

Now that we have a new key to relate step_activity to question responses we can ensure it works.

This completes iteration 1. We have imported the data, completed basic data quality checks and completed any feature engineering challenges that have presented themselves. We have also taken the identified keys and proved merging of data frames in preparation for iteration 2.

Iteration 2

Business Understanding - Hypothesis

Data Understanding

Data Preparation

Conclusions

References

Data Science Alliance, 2021, What is CRISP-DM, https://www.datascience-pm.com/crisp-dm-2/, Accessed: 26/11/2021

We can see that there are far more completed steps than incomplete, and that incomplete steps are generally earlier in the course. However, there are still students who complete early steps and subsequently fail to complete later ones. There also indicates a wide spread of time to complete each step, with a variance of 103.1396. An outlier of this is *Step 2.8* which takes everyone very few days to complete.

Due to adding the stage of the course we can also see that the time to complete each unit reduces as the course wears on, as does the number of students that complete each step.

It can be seen in the barplot for the Views by Video completion that the number of views which watched the whole video dropped off as they viewer got further through. For example we can see for the first video "1.1. Welcome to the course" that 1500 views completed 5% of the video, which drops to 1200 who completed 100%. This is then presented for each video throughout the course. As the course progresses the volume of students engaging with each video reduces reasonably significantly other than "1.5 Privacy online and offline" It can also be seen that some videos see a significant drop between 95% and 100% completion.

When viewing Views by Device it can be seen that no video was viewed on a tablet or TV, all views were by Console, Desktop or Mobile. It can be seen that by far the majority of views were on a Console throughout the length of the course, so it may be concluded the videos should be designed to play best on a console. The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

By viewing Views by Location it can be seen that the majority of views are from Europe with non at all from Antarctica and few from North America, Oceania and South America. The data shown reflects the same changes in the volume of use as shown above, i.e. that "1.5 Privacy online and offline" shows an increase in views in an otherwise downward trend.

There is potential to link views with stages and we could potentially look at leavers by location and device by using this view data along with the leaver survey responses. We could also look at test performance related to video completion, location or devices.

As we can see from the weekly experience data below, across the three weeks of the course the total sentiment data provided by the student cohort went down from 73 to 45, although the experience ratings being provided remained consistent at approximately 90% providing a rating of 3.

It would be of value to understand this data as a percentage of the total patient cohort. Given the lack of a "learner_id" in this data it isn't immediately possible to relate the sentiments back to a specific learner, so this data could only be looked into by step. There may be value in understanding after which step students decided to stop providing feedback, this could be related on the "stage_id".

3 Conclusion and Hypothesis

There is a wide breath of data in the data set regarding activity throughout the course, and richness is provided by the video views data frame which includes data such as the device used, location of the student and so forth. The use of the step dataset allows us to investigate the stage of the course which events happened, and the leavers dataset allows us to look into leavers. Knowing that the location work has been completed previously by other students, I feel that there is enough data to investigate something around student performance, stage and the decision to leave.

I hyptothesise that the lower completion of videos and lower weekly sentiment scores increase the chances of a student choosing to leave the course. This will be investigated in iteration 2 of the CRISP-DM model.