MOOC DATA ANALYSIS

Akhil Raj Rajan

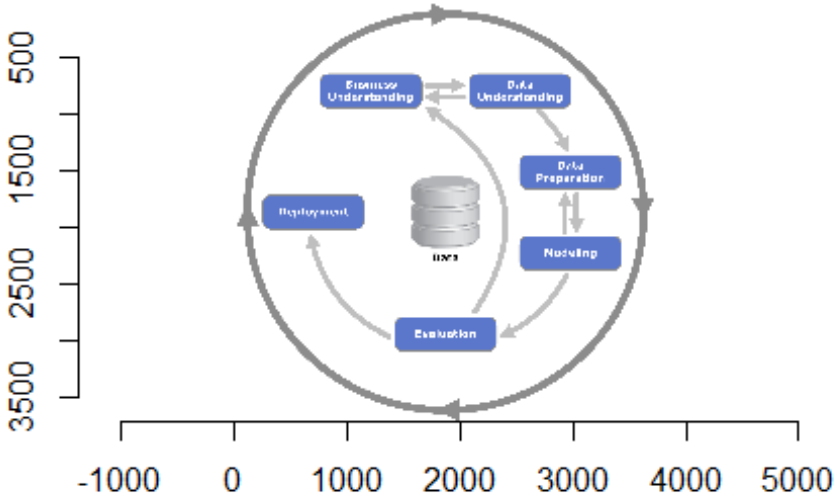
2022-11-05

INTRODUCTION

The aim of this project is to analyse the data related to massive open online course (MOOC) at Newcastle University. Basically, the data mining process has to be implemented in this case. However, the methodology/tool called CRISP-DM(Cross Industry Standard Process for data mining) can help to explore the data and provide some useful information from the raw data. CRISP-Dm is the commonly used data mining method in such cases worldwide. It consists mainly of six steps, and the process flow is not one-directional that is depending on the situation some steps will be executed multiple times. This idea is clearly explained in the below graph.

The stages of CRISP-DM and the data mining process are shown below. In this figure, a single arrow shows a phase's dependence on the other, and a double arrow implies a repeated process.

Business understanding, Data Understanding, Modelling, Evaluation, and Deployment are the six phases of CRISP-DM.



Business Understanding

The first phase of the CRISP-DM technique is business understanding or domain understanding. By analysing, processing, and putting various algorithms into use, determine the sector of business that will change into relevant information at this point. Understanding business helps you identify problems, available resources (both hardware and people), and goals.

Data Understanding

The second stage is closely related to the first understanding. Here, the meaning of the data in the real world is identified and conveyed to the stakeholders properly. The idea is that the person working on this data should understand the importance and relations of each variable in the data before proceeding to data mining.

E.g., Primary keys, foreign keys, data type, number of observations and meaning of each field in business

Data Preparation

In this stage, Data is prepared for analysis. Raw data collected from the source cannot be used directly, instead, it has to be prepared for starting the data mining process. In most cases, the following steps are carried out.

Data cleaning- e.g., removing unwanted characters like space from the data. This step is very important because it significantly affects the joining and comparison of the data since this space may not be visible to the human eye at first.

Data type changing- This step is also an important part, for example, if there is a column with string data type but on checking only numeric values are found, then changing the data type of the field to numeric will increase the speed of accessing the data. Additionally, some inbuilt functions are specific to certain data types alone, so using the original data type fails in leveraging these functions.

Model

This is the core stage in CRISP-DM which create a model that replicates the real-world situation to solve the problem. Different approaches can be used for solving a problem. Some of them are

- Simple Exploratory analysis
- Principle Component Analysis
- Linear Discriminant Analysis
- Regression model
- Decision tree
- Random forest

Some times multiple models is used for solving a problem in CRISP-DM

Evaluation

In this phase, different models are compared and ranked based on their performance, algorithmic simplicity and deployment cost. The best model is selected based on these criteria. This section includes suggestions, recommendations and criticisms.

Deployment

The deployment stage can range in complexity from deploying a repeatable data mining process across the company to something as simple as producing a report. This depends on the company's goals. Many times, the customer executes the deployment stages rather than the data analyst. However, even if the deployment effort is handled by the analyst, it is crucial for the customer to know upfront what steps must be taken.

The deployment phase consists of mainly our tasks. They are

- making preparations for deployment
- organizing, observing, and maintenance
- presenting final findings
- reviewing the outcomes

BUSINESS UNDERSTANDING

The first phase of the CRISP-DM technique is business understanding. The important stage in this phase is understanding the business requirement and problems. The objective of this data mining is well established and generates a good plan for the exploration in this step.

DATA UNDERSTANDING

A massive open online course (MOOC) at Newcastle University named "Cyber Security: Safety At Home, Online, and in Life" has been offered seven times to the general public using the online education platform FutureLearn.

- learner_id – The individual ID assigned to each student enrolled in the course.
- enrolled_at – This column contains the learner's enrollment date in the course.
- unrolled_at – The date the learner dropped out of the programme.
- role – defines the position of the registered candidate.
- fully_participated_at - Indicates the time the learner finished the course..
- purchased_statement_at – The purchase statement's generation date is indicated in this column.
- gender – displays the learner's gender who has signed up for the course.
- country – Indicates the country the learner selected while registering for the course.
- age_range – shows the range of registered students' ages.

- highest_educational_level - A learner's level of education.
- employment_status – specifies the learner's job status
- employment_area - Indicates the field in which the learned are employed
- detected_country – This field contains information about the learner's country as determined by their network address.

There are 37296 rows in this data, It is not a big number so normal procedures can be followed. The dimensions like employment area and detected country can be the foreign keys to some other tables.

The primary key for this data is learners_id as it can identify unique observations alone.

A brief Idea about the data can be seen in the below table.

Column Name	Data Type	Example data
learner_id	chr	160d6600-ea0e-4568-bfa9-5d7cd5b8e61b
enrolled_dttm	POSIXct	2016-08-10 14:28:49
unenrolled_dttm	POSIXct	2016-08-10 14:28:49
role	chr	learner
fully_participated_dttm	POSIXct	2016-08-10 14:28:49
purchased_statement_dttm	POSIXct	2016-08-10 14:28:49
sex	chr	male
country	chr	GB
age_range	chr	46-55
highest_education_level	chr	university_degree
employment_status	chr	working_part_time
employment_area	chr	teaching_and_education
detected_country	chr	GB
run	num	2

DATA PREPARATION

1.Appended all the 7 data to a single data frame after adding a column to indicate the run for easy processing. This will helps to anlayse the data based on run in future.

```
#loading the all the data to a data frame

cyber_s_e_1=cyber.security.1_enrolments
cyber_s_e_2=cyber.security.2_enrolments
cyber_s_e_3=cyber.security.3_enrolments
cyber_s_e_4=cyber.security.4_enrolments
cyber_s_e_5=cyber.security.5_enrolments
cyber_s_e_6=cyber.security.6_enrolments
cyber_s_e_7=cyber.security.7_enrolments
```

```
#adding a column "year" to all the data frames for representing the year.

cyber_s_e_1=cyber_s_e_1 %>% mutate(run=1)
cyber_s_e_2=cyber_s_e_2 %>% mutate(run=2)
cyber_s_e_3=cyber_s_e_3 %>% mutate(run=3)
cyber_s_e_4=cyber_s_e_4 %>% mutate(run=4)
cyber_s_e_5=cyber_s_e_5 %>% mutate(run=5)
cyber_s_e_6=cyber_s_e_6 %>% mutate(run=6)
cyber_s_e_7=cyber_s_e_7 %>% mutate(run=7)
cyber_s_e_all=rbind(cyber_s_e_1,cyber_s_e_2,cyber_s_e_3,cyber_s_e_4,cyber_s_e
_5,cyber_s_e_6,cyber_s_e_7)
```

2.Checked for duplicates and no duplicates were found

Removing duplicates is necessary because it may give false information about the data.

```
count(unique(cyber_s_e_all))

## # A tibble: 1 × 1
##       n
##   <int>
## 1 37296

count(cyber_s_e_all)

## # A tibble: 1 × 1
##       n
##   <int>
## 1 37296

#if the values are equal then there are no duplicates
```

3.Renamed table names to the most appropriate names.

This will help the stakeholders to get an overall idea of the data just by viewing it. This will also help the new analysts to identify the datatypes as well.

```
#renaming table names
names(cyber_s_e_all)[names(cyber_s_e_all) == 'enrolled_at'] <-
'enrolled_dttm'
names(cyber_s_e_all)[names(cyber_s_e_all) == 'unenrolled_at'] <-
'unenrolled_dttm'
names(cyber_s_e_all)[names(cyber_s_e_all) == 'fully_participated_at'] <-
'fully_participated_dttm'
names(cyber_s_e_all)[names(cyber_s_e_all) == 'purchased_statement_at'] <-
'purchased_statement_dttm'
names(cyber_s_e_all)[names(cyber_s_e_all) == 'gender'] <- 'sex'
```

4.Data cleaning by removing unwanted spaces and removing ambiguous values

```
#data cleaning by removing unwanted spaces and removing ambiguous values

cyber_s_e_all$unenrolled_dttm <- trimws(cyber_s_e_all$unenrolled_dttm, which
= c("both"))
cyber_s_e_all$learner_id <- trimws(cyber_s_e_all$learner_id, which =
c("both"))
cyber_s_e_all$role <- trimws(cyber_s_e_all$role, which = c("both"))
cyber_s_e_all$sex <- trimws(cyber_s_e_all$sex, which = c("both"))
cyber_s_e_all$country <- trimws(cyber_s_e_all$country, which = c("both"))
cyber_s_e_all$age_range <- trimws(cyber_s_e_all$age_range, which = c("both"))
cyber_s_e_all$highest_education_level <-
trimws(cyber_s_e_all$highest_education_level, which = c("both"))
cyber_s_e_all$employment_status <- trimws(cyber_s_e_all$employment_status,
which = c("both"))
cyber_s_e_all$employment_area <- trimws(cyber_s_e_all$employment_area, which
= c("both"))
cyber_s_e_all$detected_country <- trimws(cyber_s_e_all$detected_country,
which = c("both"))

cyber_s_e_all["unenrolled_dttm"][cyber_s_e_all["unenrolled_dttm"] == ""] <-
"9999-12-31 00:00:00 UTC"
cyber_s_e_all["fully_participated_dttm"][cyber_s_e_all["fully_participated_dt
tm"] == ""] <- "9999-12-31 00:00:00 UTC"
cyber_s_e_all["purchased_statement_dttm"][cyber_s_e_all["purchased_statement_
dttm"] == ""] <- "9999-12-31 00:00:00 UTC"
```

5.Data types were modified appropriately

This will help the data analysis in future to use some inbuilt functions specific to each data type. For example, extracting the year value from enrolled_dttm will be easy using the function format().

```
#data type changing
cyber_s_e_all$run<-as.numeric(cyber_s_e_all$run)
cyber_s_e_all$enrolled_dttm<- as.POSIXct( cyber_s_e_all$enrolled_dttm, tz =
"UTC" )
cyber_s_e_all$unenrolled_dttm<- as.POSIXct( cyber_s_e_all$unenrolled_dttm, tz
= "UTC" )
cyber_s_e_all$fully_participated_dttm<- as.POSIXct(
cyber_s_e_all$fully_participated_dttm, tz = "UTC" )
cyber_s_e_all$purchased_statement_dttm<- as.POSIXct(
cyber_s_e_all$purchased_statement_dttm, tz = "UTC" )
```

MODELLING

The objective of this analysis is to find the answers to the business question using CRISP-DM model. For answering the first question, a histogram plot was generated for the different age groups that show the frequency of enrolled candidates for each group. Then, the majority of the age group is found using a bar plot since it is the best visual method for this task. The procedure is given below.

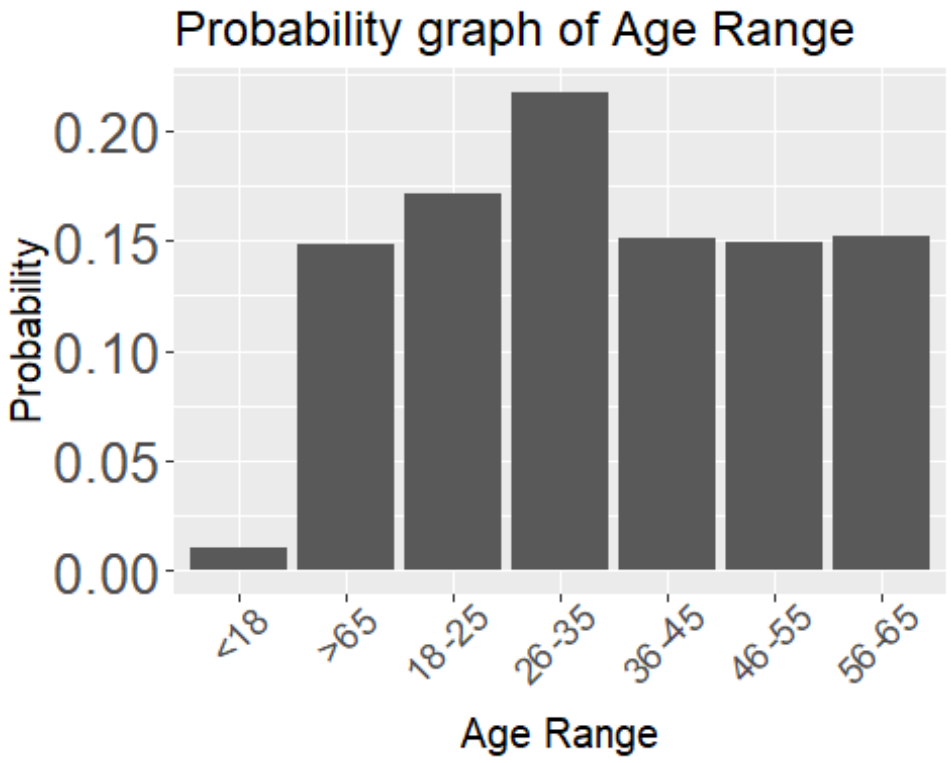
age_range= Age group, n= number of candidates under this category, prob_n= n/Total

1.age_range

The records with unknown as the age_limit were eliminated for getting accurate model

age_range	n	prob_n
<18	42	0.0104270
>65	598	0.1484608
18-25	691	0.1715492
26-35	875	0.2172294
36-45	608	0.1509434
46-55	602	0.1494538
56-65	612	0.1519364

Plotting a probability diagram for this data



Comments: The above graph shows that the age group “26-35” is the majority enrolled group. Additionally, the age groups 36-45,46-55,56-65 all seem to be uniformly distributed.

Evaluation

- The model shows that the majority of candidates are from the age group 26-35. They can be categorized as youth people.
- For the remaining age group above 35, the total number of candidates is approximately the same in each group.
- However, only a few candidates are enrolled under the age of 18.
- The probability distribution of the data can be seen as more aligned to uniform distribution except for the age group below 18.
- The overall data spread across all the age groups greater than 18.

BUSINESS UNDERSTANDING-CYCLE 2

For further exploration, the employment status of these candidates under the age 26-35 is analysed and its result gives the idea about how these candidates under this age group relate to the job type. The answer to this question gives a useful insight into data mining for business requirements.

DATA UNDERSTANDING

Here, we are using the same data as that in the first cycle. So, additional information about the data is not required.

DATA PREPARATION

Since there are some “unknown” values in the employment status field, the rows with unknowns are filtered out based on the assumption that it is invalid.

MODELLING

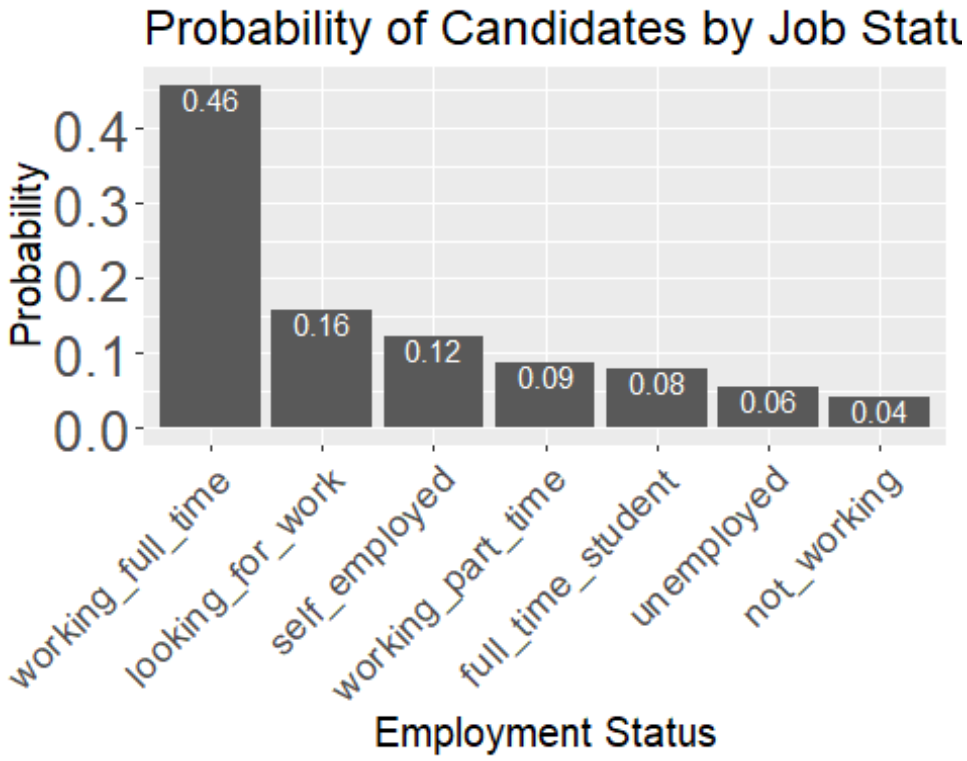
The first model showed that the majority of the candidates are from the age group 26-35. Based on this output from cycle 2, the previous data is filtered and provided as input for cycle -2. Here, the main focus is on answering the next question that is “what is the employment status of the majority age group?”. To answer, this question, a histogram plot showing the employment status of the age group 26-35 is used. The process is given in the below steps.

employment_status= employment status, n= number of candidates under this category, prob_n= n/Total

employment_status	n	prob_n
not_working	35	0.0404624

employment_status	n	prob_n
unemployed	48	0.0554913
full_time_student	69	0.0797688
working_part_time	75	0.0867052
self_employed	106	0.1225434
looking_for_work	136	0.1572254
working_full_time	396	0.4578035

Plotting a probability diagram regarding the candidate under the age group 26-35 for each employment status.



comments: The above-given figure show that the candidate under 26-35 (majority age group) are full-time time employees. Also, The candidates who are looking for a job stand second higher in number under the same category. lastly, The number of learners who are not working attending the course under this category is less in number.

Evaluation

- The majority of candidates are working as a full-time employees under the age group 26-35. (0.46%)
- The number of candidates who are not working under this group is relatively very low (4%)

- This makes sense that people who are working full time might need to up their skills which resulted in this variation in the data.

BUSINESS UNDERSTANDING-CYCLE 3

So far, the information about the age group and the employment status is analysed for the enrolled candidates. However, a year-wise analysis will give better insight into the decision making like what measures should be taken to increase the profit. should it be urgent? etc.

To answer the above question, an year-wise analysis is used.

DATA UNDERSTANDING

Here, we are using the same data as that in the first cycle. but a new field is added which shows the date each of the candidates is enrolled.

DATA PREPARATION

The same data is used in the cycle too. But, from the enrollment_dttm field year is extracted and utilized for analysis. The code for this is shown below

```
#Students enrolled  with year

cyber_s_e_all_yr =cyber_s_e_all %>% filter(age_range=="26-35" &
employment_status == "working_full_time") %>%
  mutate(
    enrolled_year =format(enrolled_dttm, format="%Y"),
  )

# The data frame contain the year of allotment of all candidate
```

Modelling cycle-3

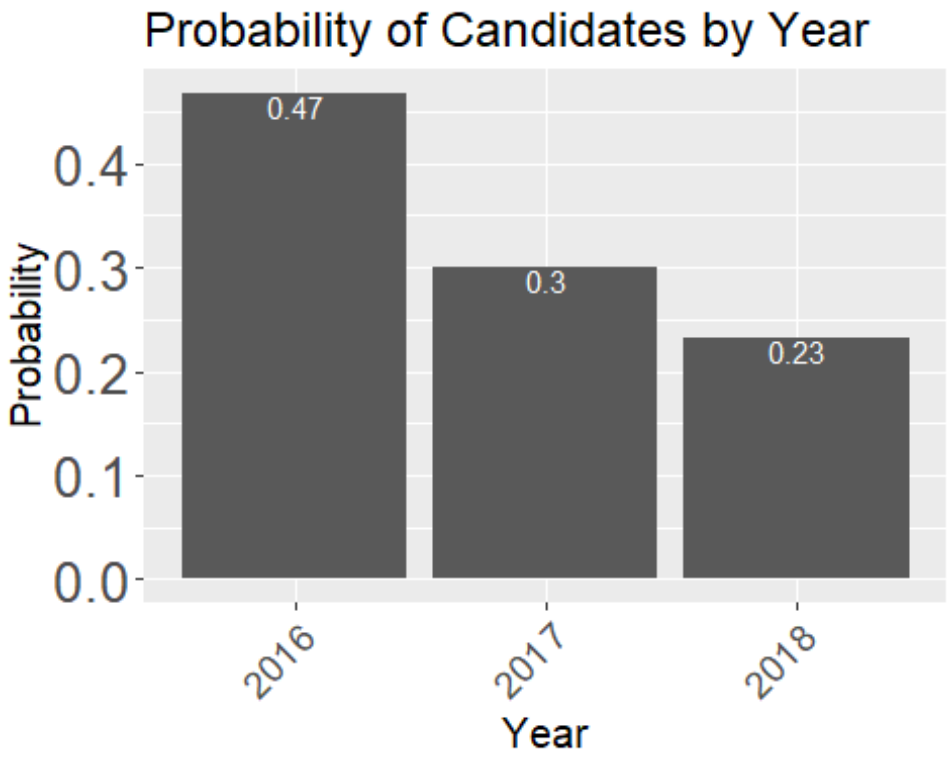
Here, The input for this cycle is the output from the previous one, and the data is compared and analysed over different years. . Now, the enrolled_year can be used for the year-wise analysis of the candidate under the age category 26-35 and working full time.

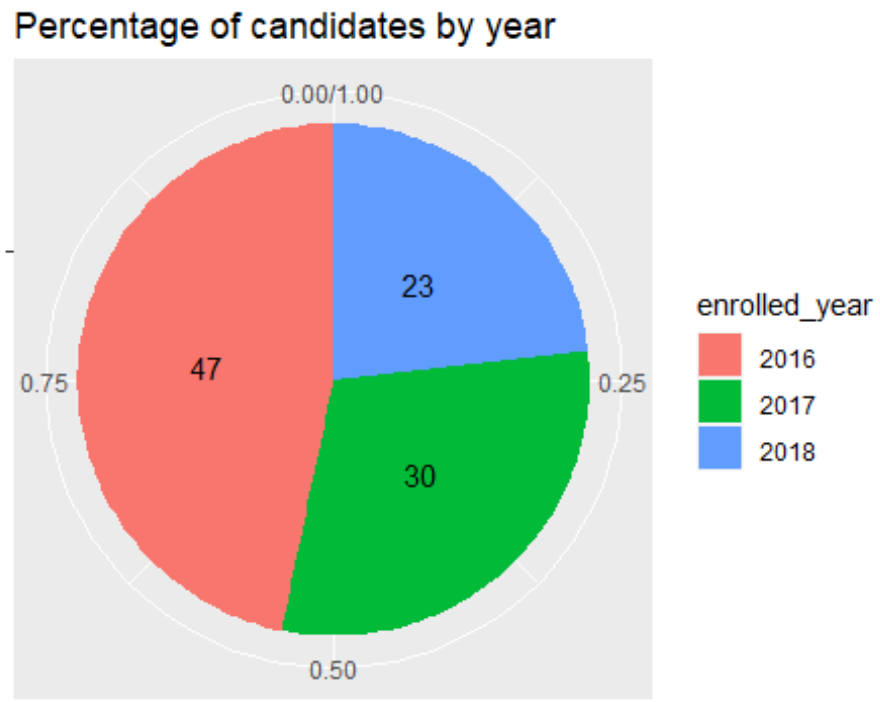
The table shows the percentage of candidates under the age category 26-35 and working full-time for each year.

enrolled_year= Year enrolled, n= number of candidates under this category, prob_n= n/Total

enrolled_year	n	prob_n
2018	92	0.2323232
2017	119	0.3005051
2016	185	0.4671717

The bar plot and Pie chart of the above table are given below for better understanding





Comments: The number of candidates under the age category 26-35 who are working full time is decreasing from 2016 to 2018

Evaluation

The result of three CRISP-DM cycles is discussed in this section. The questions that need to be answered are:

1. which age group were enrolled more compared to others?
2. What is the employment status of the majority age group who joined in this online learning platform?
3. How does the number of enrolled candidates varies over time given that the candidates are under the 26-35 age group and working as a full-time employee?

The first two parts were already answered, and the final result is that the number of enrolled candidates under this mentioned condition is decreasing from 2016 to 2018 which is not good for this organisation. This can happen due to many reasons, but the organisation need to take immediate decisions to reverse this graph.

Deployment

With the help of the first two cycles, the number /percentage of candidates who are working as full-time employees under the age group 26-35 which is also the majority age group is found. The last cycle deals with the year-wise analysis of the output of the second cycle which indicates the number of candidates is decreasing in each year from 2016 to 2018.

The importance of this exploratory analysis is that it shows some insights into the existing trends of this online learning platform regarding its business growth. The output of this design method can be read by a business analyst or any stakeholders and they can take immediate actions.

The graphical reports generated by these models can be easily reported using any data visualization tools like Power BI, Tableau etc. Since the method follows an iteration or cycle-like strategy new information with more clarity than the previous one is obtained which is one of the advantages of the crisp-dm methodology. This also enables the stakeholder to initiate the steps from scratch and improve their explorations, data mining, analysis and exploration phases immediately as possible since CRISP-DM stands as the base for all this process.

Here, in this example, the full potential CRISP-DM is not used since no machine learning models were implemented. Instead if ML algorithms like decision tree, Random forest are used we can predict some of the responsive variables effectively, and with each cycle, the prediction accuracy will be improved.