CSC8631 Report

Morgan Frodsham

08/11/2021

FutureLearn Learning Analytics for Newcastle University

Business understanding

Background

This report details the process and findings of exploratory data analysis for Newcastle University's FutureLearn course called "Cyber Security: Safety at Home, Online, in Life." Data analysis of educational platforms, such as FutureLearn, is a key part of learning analytics, which is defined as *"the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimising learning and the environments in which it occurs." 1st International Conference on Learning Analytics and Knowledge* (2011) Learning analytics has the potential to help educational institutions undertake data-driven decision-making Long and Siemens (2011) that better support students through their educational journey Shacklock (2016).

Business objectives and success criteria

FutureLearn is a massive open online course (MOOC) provider. It promotes itself as *"a powerful new way to learn online"* FutureLearn (n.d.), designed with the principles of effective learning. Learning effectiveness can be indicated by student engagement, their success rate, and completion time S. Hubalovsky and Musilek (2018). A core objective of FutureLearn is to improve learning effectiveness.

The objective of this data analysis is to better understand the learning effectiveness of Newcastle University's Future Learn Cyber Security course. The process and findings outlined by this report explore the question: "Do learners have a higher chance of success the more data they share with FutureLearn?" Here, "a higher chance of success" means that learners are more likely to complete the course. The data shared by learners refers to the optional data requested by FutureLearn; specifically, the provision of personal identifiers like demographic information. The provision of this data is taken to be an indicator of student engagement with the course. For this analysis the business success criteria is to give useful insights into the relationship between learner engagement and learner success rate.

(objectives that were considered and rejected, and rational for why?)

Project plan

The deadline for this project is 16:30, 3 December 2021. (This section lists the stages to be executed in the project, together with their duration, resources required, inputs, outputs, and dependencies. Where possible, it should make explicit the large-scale iterations in the data mining process—for example, repetitions of the modeling and evaluation phases)

Initial assessment of tools and techniques

(methodology, This section gives an initial view of what tools and techniques are likely to be used and how. It describes the requirements for tools and techniques, lists available tools and techniques, and matches them to requirements.)

Inventory of resources

Newcastle University has provided x number of data sets for exploration.

Tidyverse for R is the library used for the analysis, ggplot2 for R is the library used for visualisations, ProjectTemplate for R is used to improve the reproducibility of this analysis, Git is used for version control, and all written documentation is created through RMarkdown for R.

Requirements, assumptions and constraints

This project, as well as the process and findings outlined in this report, is limited to the data analysis pipeline. (assumptions made about the nature of the problem, constraints imposed on the project (technical suite)?) correlation, confounding variables

Risks and contingencies

Risk re un-reproducibility, version control, etc. - check data flow lecture) The technical suite used aims to apply good programming practice, and reduce common risks that might delay the project or cause it to fail.

Costs and benefits

This section describes the costs of the project and predicted business benefits if the project is successful (e.g., return on investment). Other less tangible benefits (e.g., customer satisfaction) should also be highlighted.

Terminology

tibble = table

Data mining goals and success criteria

The exploratory data analysis outlined by this report explores the question: “Do learners have a higher chance of success the more data they share with FutureLearn?” To understand this, the data mining aims to determine:

1. What is the gross success rate of learners? The data mining success criteria is the percentage of all learners who complete the course.
2. How many learners provide any demographic data? The data mining success criteria is the percentage of all learners who provide optional demographic data such as gender, age range or employment status.
3. What is the success rate of learners who provide demographic data? The data mining success criteria is the percentage of learners who complete the course and provide demographic data.
4. What is the success rate of learners who do not provide demographic data? The data mining success criteria is the percentage of learners who complete the course and do not provide any demographic data.

5. Is there any correlation between success and the provision of demographic data? The data mining success criteria is the difference between (and standard deviation of?) each of the aforementioned learner success rates.

Data understanding

Initial data collection

Newcastle University's FutureLearn cyber security course has run seven times. The first run provided six different databases, the second run provided seven, and the others provided eight. There are 53 different databases for potential exploration.

It is unclear whether there have been any problems in data acquisition or extraction as Newcastle University has provided the data.

Data description

The first run of Newcastle University's FutureLearn cyber security course provides the following databases: - archetype survey responses (id, learner_id, responded_at, archetype); - enrolments (learner_id, enrolled_at, unenrolled_at role, fully_participated_at, purchased_statement_at gender, country age_range, highest_education_level, employment_status, employment_area detected_country); - leaving survey responses (id, learner_id, left_at leaving_reason, last_completed_step_at, last_completed_step, last_completed_week_number, last_completed_step_number); - question responses (learner_id, quiz_question, question_type, week_number, step_number, question_number, response, cloze_response, submitted_at, correct); - step activity (learner_id step, week_number step_number, first_visited_at, last_completed_at); and - weekly sentiment survey (id responded_at, week_number, experience_rating, reason).

The second run provides the same databases as well as a databases specifying the course team members (id, first_name, last_name, team_role, user_role).

The other five runs provide the same databases as the second run as well as a database on the course's video stats (step_position, title video_duration, total_views, total_downloads, total_caption_views, total_transcript_views, viewed_hd, viewed_five_percent, viewed_ten_percent, viewed_twentyfive_percent, viewed_fifty_percent, viewed_seventyfive_percent, viewed_ninetyfive_percent, viewed_onehundred_percent, console_device_percentage, desktop_device_percentage, mobile_device_percentage, tv_device_percentage, tablet_device_percentage, unknown_device_percentage, europe_views_percentage, oceania_views_percentage, asia_views_percentage, north_america_views_percentage, south_america_views_percentage, africa_views_percentage, antarctica_views_percentage).

Data quality

(Completeness/accuracy - not completely complete) p The data collected by the first run of the course is relatively sparse compared to the seventh run of the course. In runs three and four more data is being collected; this could be because there are fewer optional sections (such as the archetype or leaving survey), but it is unknown why this change has occurred.

Data exploration

To address the business objective and success criteria, the exploratory analysis outlined by this report begins with the enrolments database because it includes the field fully_participated_at, which is considered to mean course completion, as well as learners' demographic data.

Goal 1 (Gross Success Rate)

To understand the success rate of all learners who undertake the course, the total number of learners who have undertaken the course across all seven runs needs to be determined. All seven runs were combined, two filters were created to show how many learners completed (complete) and did not complete the course (incomplete).

```
# Calculate the success rate of all learners who enrolled in the course.
total_learners <- (2154 + 35142) %>% #How many total learners are there?
  print()
```

```
## [1] 37296
```

```
gross_success_rate <- (100*(2154/total_learners)) %>% #Success rate as a percentage
  round(., digits = 2) %>% #Success rate rounded to two digits
  print()
```

```
## [1] 5.78
```

This means that there are 37,296 learners who have enrolled in Newcastle University's FutureLearn cyber security course, and 5.78% of them have completed it (the gross success rate).

Goal 2 (Demographic Data)

Demographic data refers to the following fields that learners could provide: - gender, - country, - age_range, - highest_education_level, - employment_status, and - employment_area.

To investigate whether learners had provided any of this demographic data, a new data set called 'FL1' was created with new columns (that corresponded to the aforementioned demographic data) showing TRUE or FALSE for whether the learner had shared something.

```
# Create a new column called "count" in FL1 containing the value (count of columns) where there is a a
true_counts <- FL1 %>%
  mutate(num_true = rowSums(.[ids_of_declared_cols] != FALSE)) # Count up how many of the columns have

# Count the number of demographic columns filled in by the learner who completed and did not complete t
learner_counts <- true_counts %>%
  group_by(num_true, completed) %>%
  count()

# Display a tibble of learners who have completed or not completed the course, counting number of demog
learner_counts %>%
  print(ids_of_declared_cols, true_counts, n = 15, width = Inf)
```

```
## # A tibble: 15 x 3
## # Groups:   num_true, completed [15]
##    num_true completed     n
##       <dbl> <lgl>     <int>
##  1        0 FALSE     31467
##  2        0 TRUE       1626
##  3        1 FALSE        14
##  4        2 FALSE        15
##  5        2 TRUE          4
##  6        3 FALSE        31
##  7         3 TRUE          9
##  8         4 FALSE        79
##  9         4 TRUE          7
## 10         5 FALSE       858
## 11         5 TRUE        134
## 12         6 FALSE      2676
## 13         6 TRUE        372
## 14        NA FALSE         2
## 15        NA TRUE          2
```

This table shows us how many learners who have completed (TRUE) or not completed (FALSE) the course have provided demographic data and, if so, how many demographic data fields they have provided.

There are four instances of NA in the table that should be investigated further; for now, it is assumed that these learners did not provide demographic data.

Below shows that 4,199 learners provided demographic data.

```
# Calculate the number of learners who provided demographic data
d_count <- (37296 - 31467 - 1626 - 4) %>%
  print()
```

```
## [1] 4199
```

Goal 3 (Demographic Data Success Rate)

The table above shows us how many learners provided demographic data and completed the course. This enables us to calculate the success rate of learners who have provided demographic data.

```
# Calculate how many learners provided demographic data.
d_complete_count <- (4+9+7+134+372) %>%
  print ()
```

```
## [1] 526
```

```
# Calculate the success rate of learners who provided demographic data.
d_success_rate <- (100*(as.numeric(d_complete_count))/(as.numeric(d_count))) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 12.53
```

This means that of the 4,199 learners who provided demographic data, 526 completed the course. The success rate of learners who provide demographic data is 12.53%.

Goal 4 (No Demographic Data Success Rate)

The table created as part of Goal 2 shows how many learners did not complete the course and did not provide demographic data.

```
# Calculate how many learners did not provide demographic data.
no_d_count <- (31467+1626+2+2) %>%
  print()
```

```
## [1] 33097
```

```r
# Calculate how many of those learners completed the course.
no_d_complete_count <- (1626+2) %>%
  print()
```

```
## [1] 1628
```

```r
# Calculate success rate of learners who did not provide demographic data.
no_d_success_rate <- (100*(as.numeric(no_d_complete_count))/(as.numeric(no_d_count))) %>%
  round(., digits = 2) %>%
  print()
```

```
## [1] 4.92
```

This means that of the 33,097 learners who did not provide demographic data, 1,628 learners did not complete the course. The success rate of learners who do not provide demographic data is 4.92%.

Goal 5 (Potential Correlation)

We have identified three different success rates for learners: - 5.87% Gross Success Rate - 12.53% Demographic Data Success Rate - 4.92% No Demographic Data Success Rate

It appears there may be a correlation between learners' success rate and their provision of demographic data (as a proxy for learner engagement); learners who provide demographic data are over twice as likely (2.54x) to complete the course than learners who do not provide demographic data.

Business Understanding (Second Iteration)

Business objectives and success criteria

For this analysis, the objective is to determine whether learners have a higher chance of success the more data they share with FutureLearn. An initial investigation into the data provided by Newcastle University suggests that learners do have a higher chance of success the more data they share with FutureLearn. However, the business success criteria is to give useful insights into the relationship between learner engagement and learner success rate. A better understanding of the correlation between learners' success rate and their provision of demographic data is therefore needed.

Data mining goals and success criteria

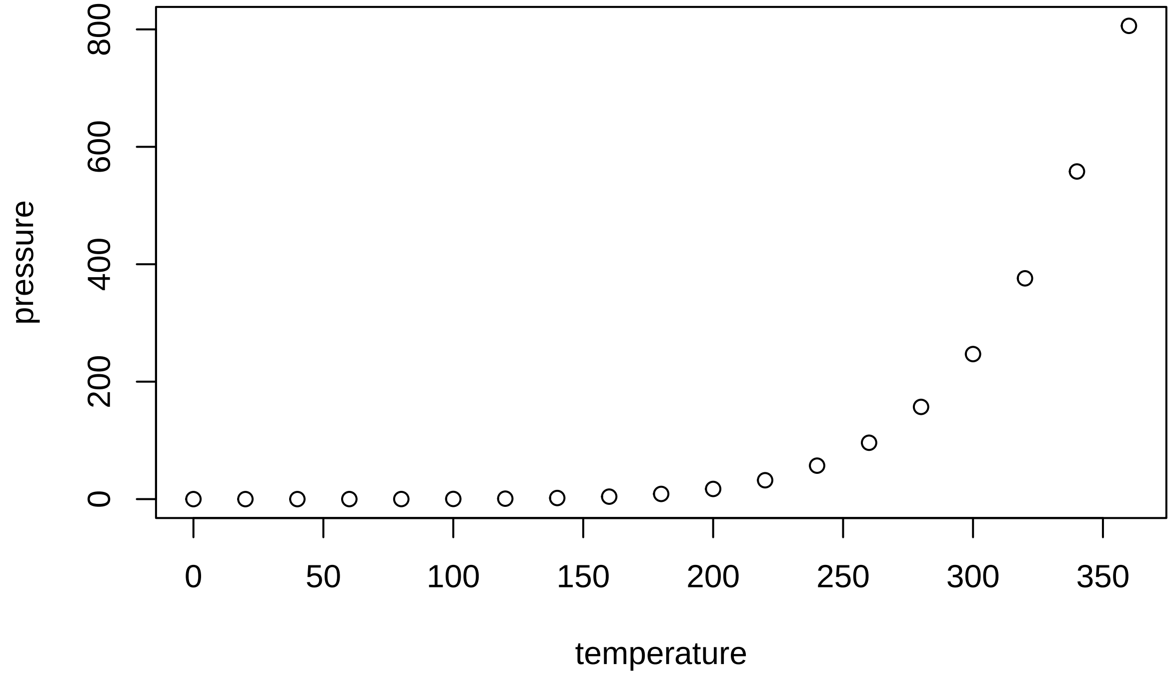
To better understand the aformentioned correlation,

Data preparation

(a description of the dataset after pre-processing and the process it was produced - basically why and how of munge stuff)

```r
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
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

Bibliography

1st International Conference on Learning Analytics and Knowledge. 2011. Banff, Alberta. https://tekri.athabascau.ca/analytics/.

FutureLearn. n.d. "The Power of Social Learning: An Effective Way to Learn." https://www.futurelearn.com/using-futurelearn/why-it-works.

Long, Phil, and George Siemens. 2011. "Penetrating the Fog: Analytics in Learning and Educaton." *EDUCAUSE Review* September/October. https://er.educause.edu/articles/2011/9/penetrating-the-fog-analytics-in-learning-and-education.

S. Hubalovsky, M. Hubalovska, and M. Musilek. 2018. "Assessment of the Influence of Adaptive e-Learning on Learning Effectiveness of Primary School Puplis." *Computers in Human Behaviour* 92. https://doi.org/10.1016/j.chb.2018.05.033.

Shacklock, Xanthe. 2016. "From Bricks to Clicks." Policy Connect; Higher Education Commission.