Recommendations arising from performing Data Analytics on FutureLearn Courses

MIGUEL BALLESTEROS

MSC DATA SCIENCE

ADRIANA WILDE

PROJECT SUPERVISOR

Insights from Data Analytics for FutureLearn

MIGUEL BALLESTEROS

MSC DATA SCIENCE

ADRIANA WILDE

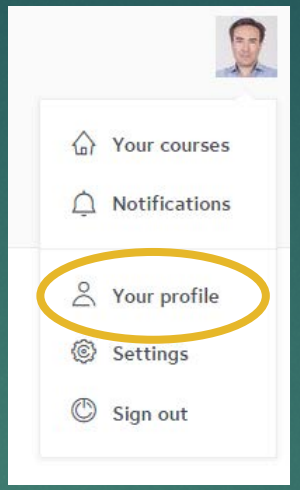
PROJECT SUPERVISOR

Agenda

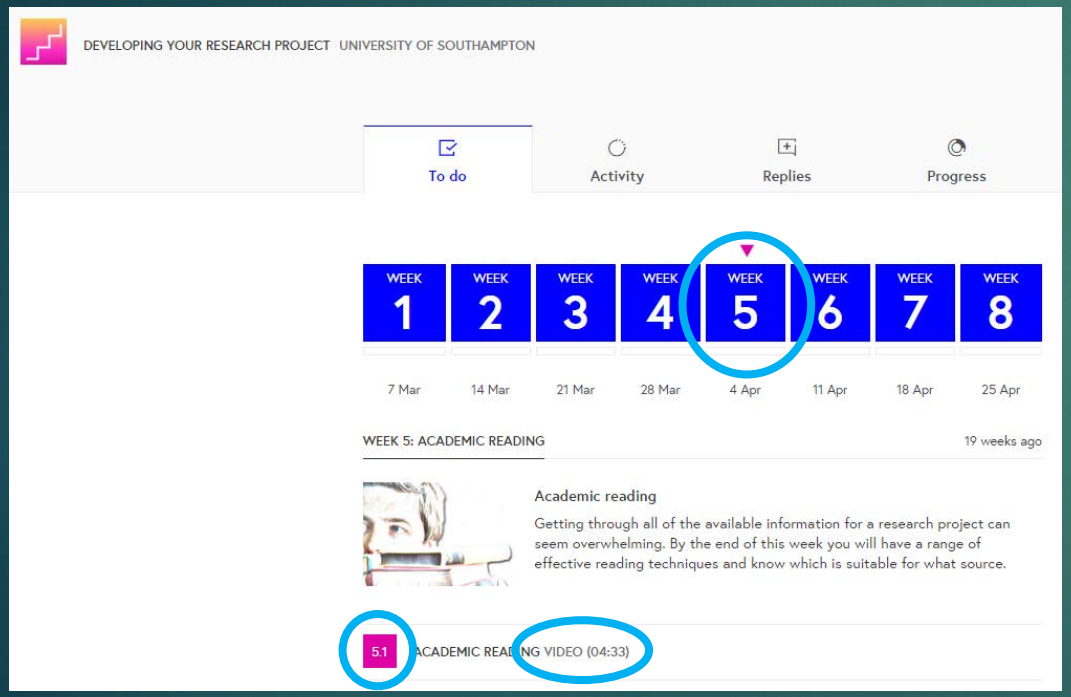
- Scenario
- Data
- Analysis
  - Learning Profiles (Visual Analytics)
  - Dropout Prediction (Machine Learning)
- Recommendations
- Future Work
- Resources

The Scenario

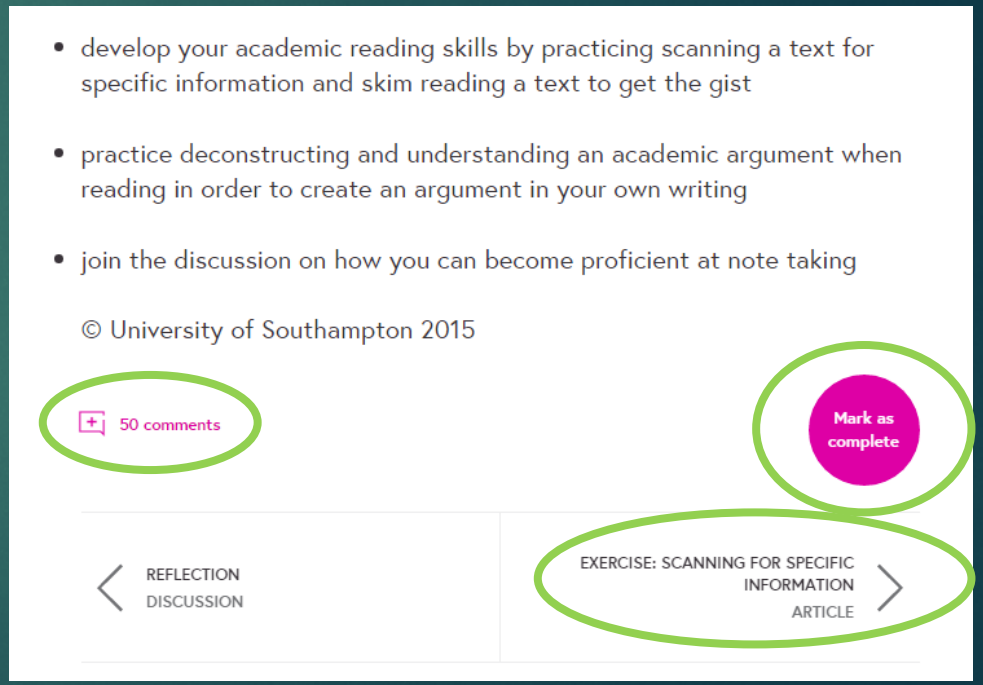
Scenario



Participants



Courses



Behaviors

Scenario

- **Learning Profiles**
- What are the key factors driving positively or negatively the participants learning experience in the online platform? and if identified, describe the ones can be considered as good or bad practices.
- What makes a good course design in terms of content variety, length and social interaction?

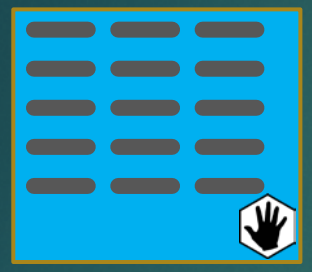
- **Predicting Dropouts**
- How many participants are likely to leave in the coming one and two weeks?

All analysis within this project use the data from 3 different courses that had multiple runs, topics and audiences. Some findings cannot be generalized!

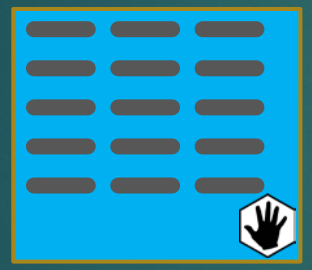
The Data

Data

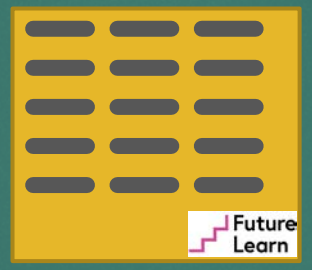
Course List



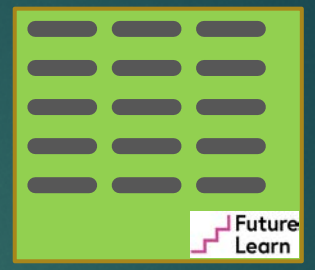
Course Details



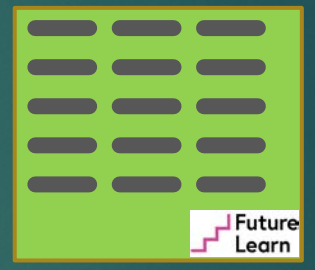
Enrolments



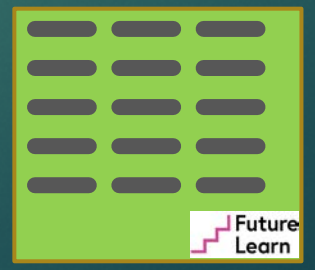
Activity



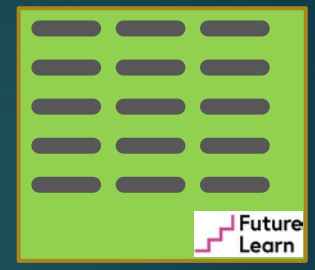
Comments



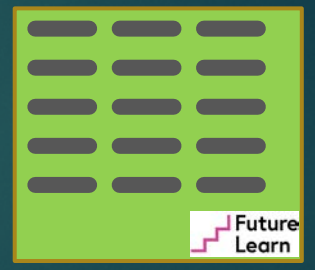
Questions



Reviews



Assignments

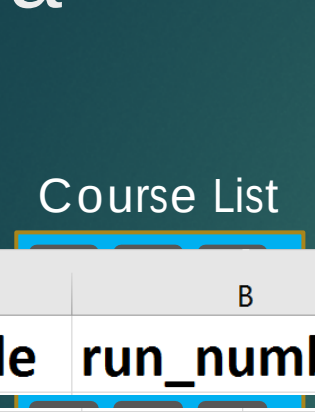


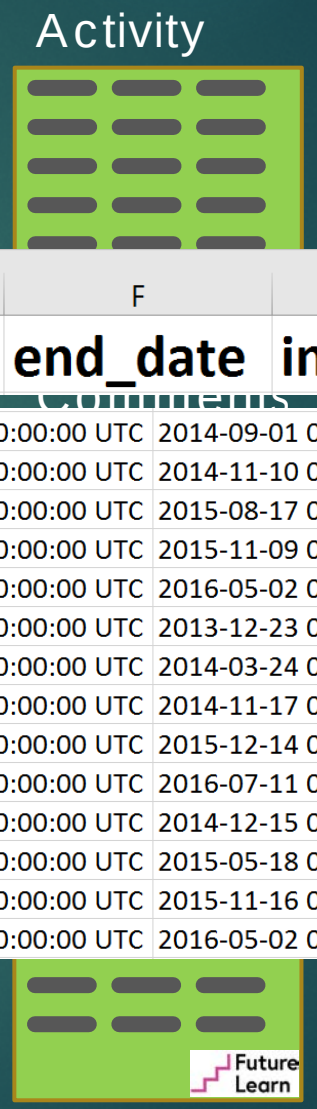
Courses

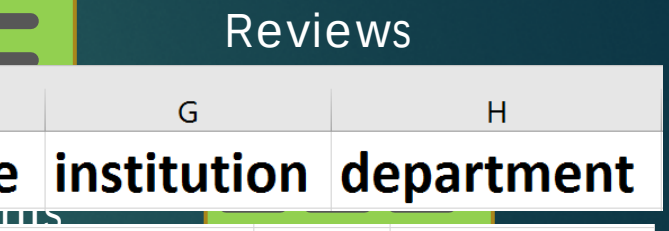
Participants

Behaviors

Data





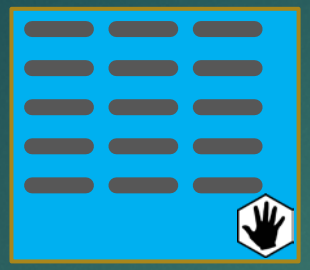


	A	B	C	D	E	F	G	H
1	short_code	run_number	short_name	full_name	start_date	end_date	institution	department
2	research-project-1	1	research-project	Developing your research project	2014-07-07 00:00:00 UTC	2014-09-01 00:00:00 UTC	UoS	Student Recruitment
3	research-project-2	2	research-project	Developing your research project	2014-09-15 00:00:00 UTC	2014-11-10 00:00:00 UTC	UoS	Student Recruitment
4	research-project-3	3	research-project	Developing your research project	2015-06-22 00:00:00 UTC	2015-08-17 00:00:00 UTC	UoS	Student Recruitment
5	research-project-4	4	research-project	Developing your research project	2015-09-14 00:00:00 UTC	2015-11-09 00:00:00 UTC	UoS	Student Recruitment
6	research-project-5	5	research-project	Developing your research project	2016-03-07 00:00:00 UTC	2016-05-02 01:00:00 UTC	UoS	Student Recruitment
7	web-science-1	1	web-science	Web Science: How the Web is changing the world	2013-11-11 00:00:00 UTC	2013-12-23 00:00:00 UTC	UoS	ECS
8	web-science-2	2	web-science	Web Science: How the Web is changing the world	2014-02-10 00:00:00 UTC	2014-03-24 00:00:00 UTC	UoS	ECS
9	web-science-3	3	web-science	Web Science: How the Web is changing the world	2014-10-06 00:00:00 UTC	2014-11-17 00:00:00 UTC	UoS	ECS
10	web-science-4	4	web-science	Web Science: How the Web is changing the world	2015-11-30 00:00:00 UTC	2015-12-14 00:00:00 UTC	UoS	ECS
11	web-science-5	5	web-science	Web Science: How the Web is changing the world	2016-06-27 00:00:00 UTC	2016-07-11 00:00:00 UTC	UoS	ECS
12	understanding-language-1	1	understanding-language	Understanding Language: Learning and Teaching	2014-11-17 00:00:00 UTC	2014-12-15 00:00:00 UTC	UoS	Languages
13	understanding-language-2	2	understanding-language	Understanding Language: Learning and Teaching	2015-04-20 00:00:00 UTC	2015-05-18 00:00:00 UTC	UoS	Languages
14	understanding-language-3	3	understanding-language	Understanding Language: Learning and Teaching	2015-10-19 00:00:00 UTC	2015-11-16 00:00:00 UTC	UoS	Languages
15	understanding-language-4	4	understanding-language	Understanding Language: Learning and Teaching	2016-04-04 00:00:00 UTC	2016-05-02 00:00:00 UTC	UoS	Languages

Courses **Participants** **Behaviors**

Data

Course List

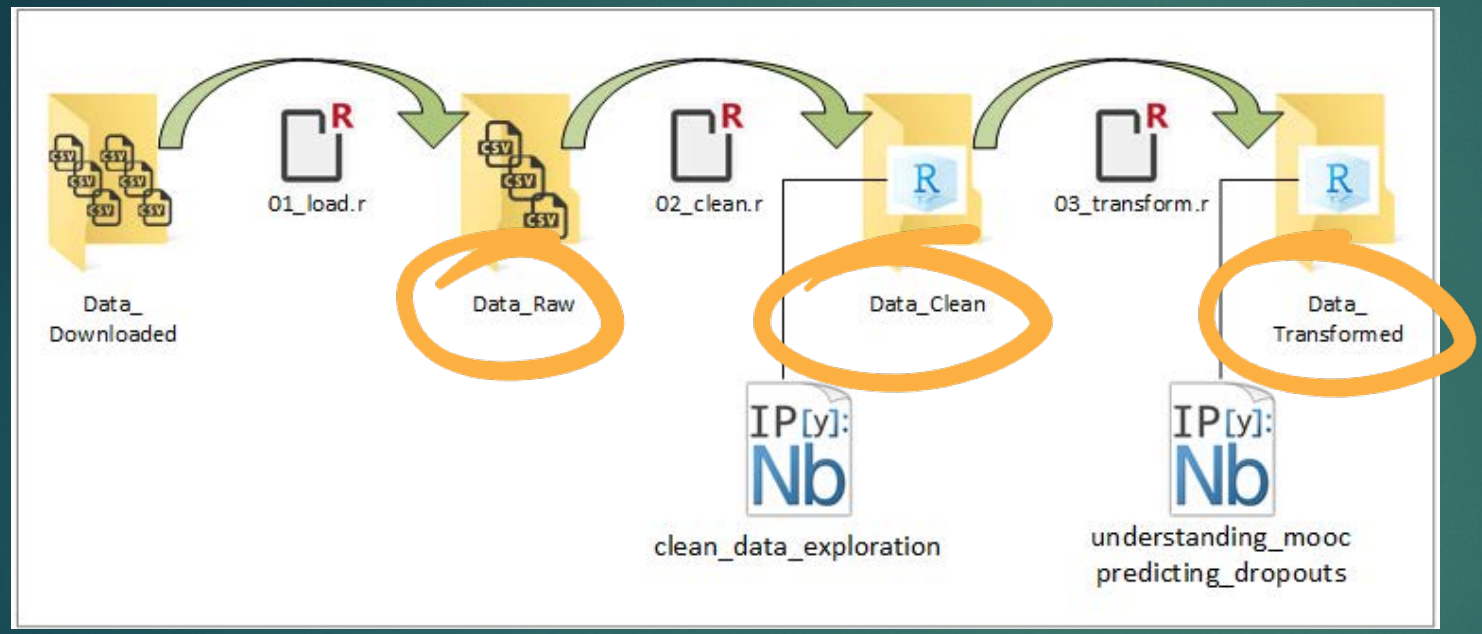


Course Details

	A	B	C	D	E	F	G	H	I
1	short_code	run_number	short_name	week_number	step_number	content_type	duration_estimated	week_start_date	week_end_date
2	research-project-1	1	research-project	1	1	video	240	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
3	research-project-1	1	research-project	1	2	article	180	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
4	research-project-1	1	research-project	1	3	discussion	300	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
5	research-project-1	1	research-project	1	4	article	420	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
6	research-project-1	1	research-project	1	5	video	240	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
7	research-project-1	1	research-project	1	6	video	240	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
8	research-project-1	1	research-project	1	7	video	240	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
9	research-project-1	1	research-project	1	8	discussion	300	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
10	research-project-1	1	research-project	1	9	article	600	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
11	research-project-1	1	research-project	1	10	article	180	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
12	research-project-1	1	research-project	1	11	discussion	300	2014-07-07 00:00:00 UTC	2014-07-14 00:00:00 UTC
13	research-project-1	1	research-project	2	1	video	240	2014-07-14 00:00:00 UTC	2014-07-21 00:00:00 UTC
14	research-project-1	1	research-project	2	2	discussion	300	2014-07-14 00:00:00 UTC	2014-07-21 00:00:00 UTC
15	research-project-1	1	research-project	2	3	video	60	2014-07-14 00:00:00 UTC	2014-07-21 00:00:00 UTC

Data

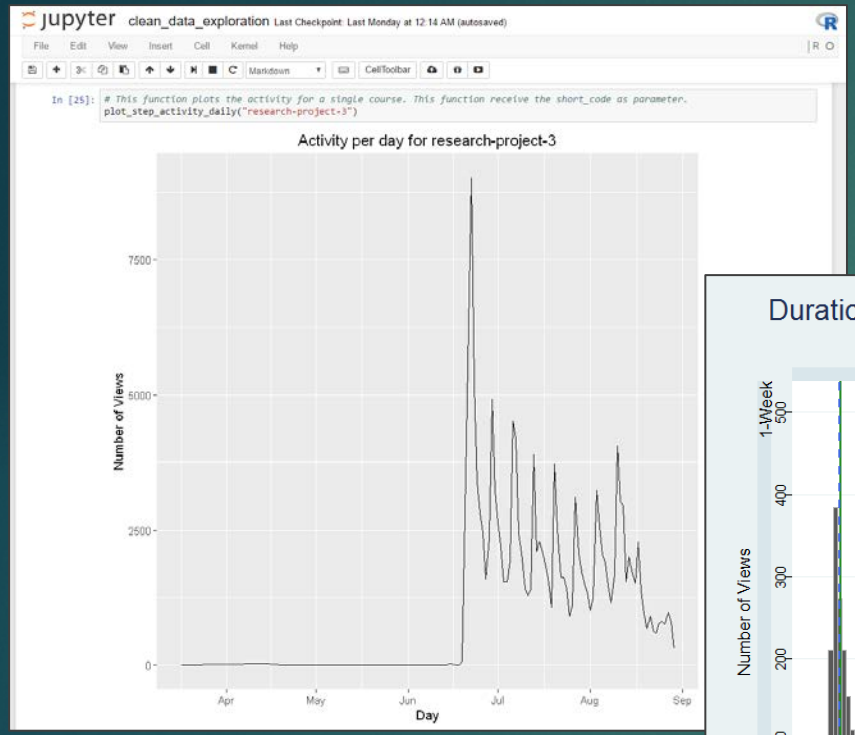
Processing Pipeline

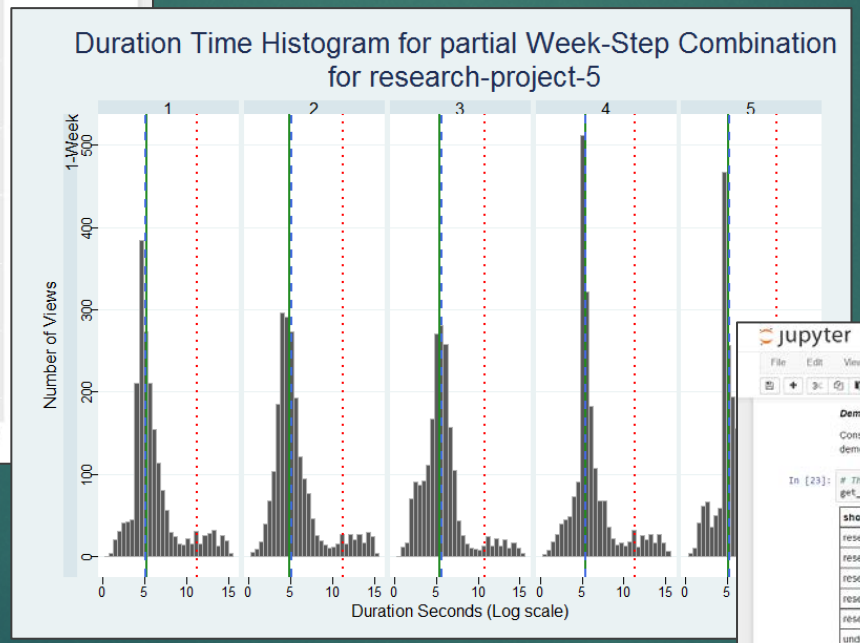


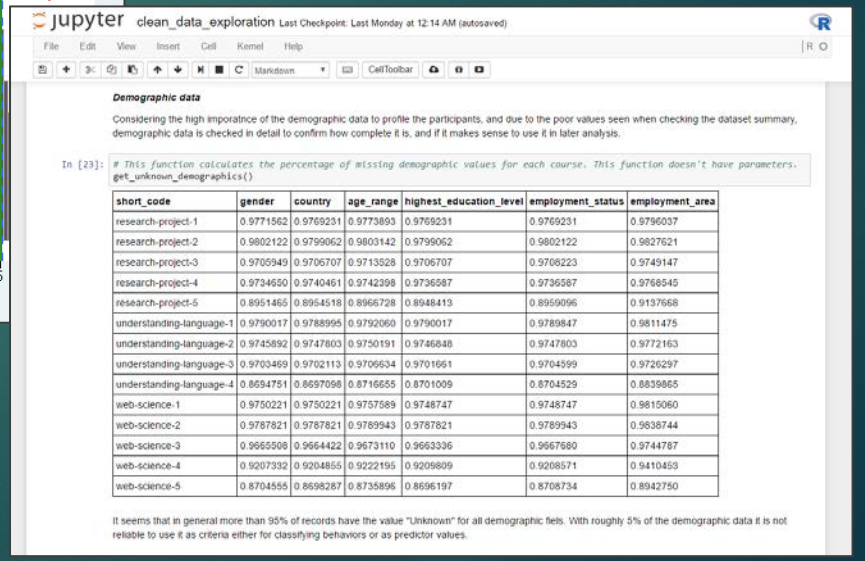
Facts!

The Analysis

Analysis – Feature Engineering

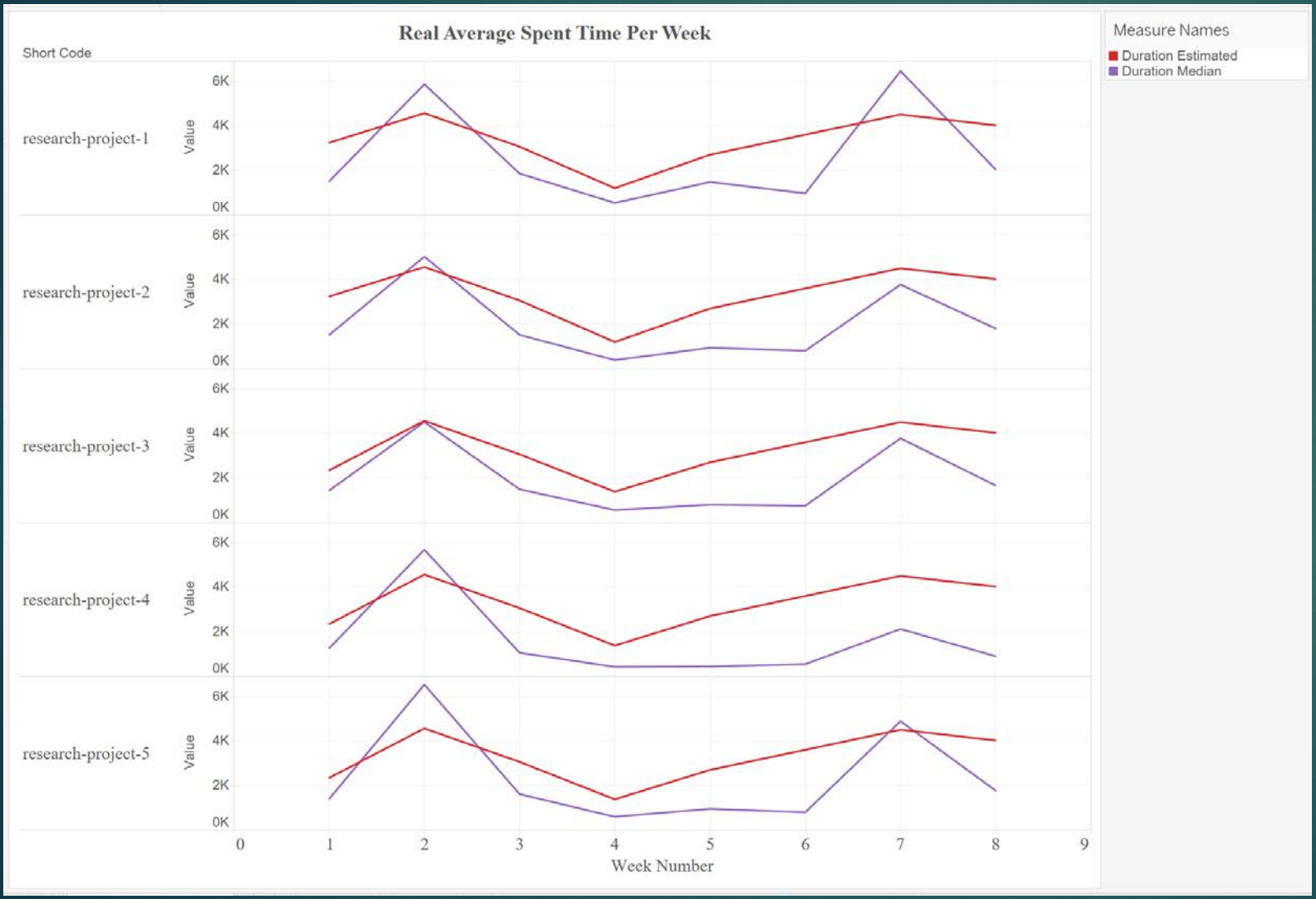






- Defined the main drivers
- Identified relevant data for transformations
- Removed irrelevant and distorting data

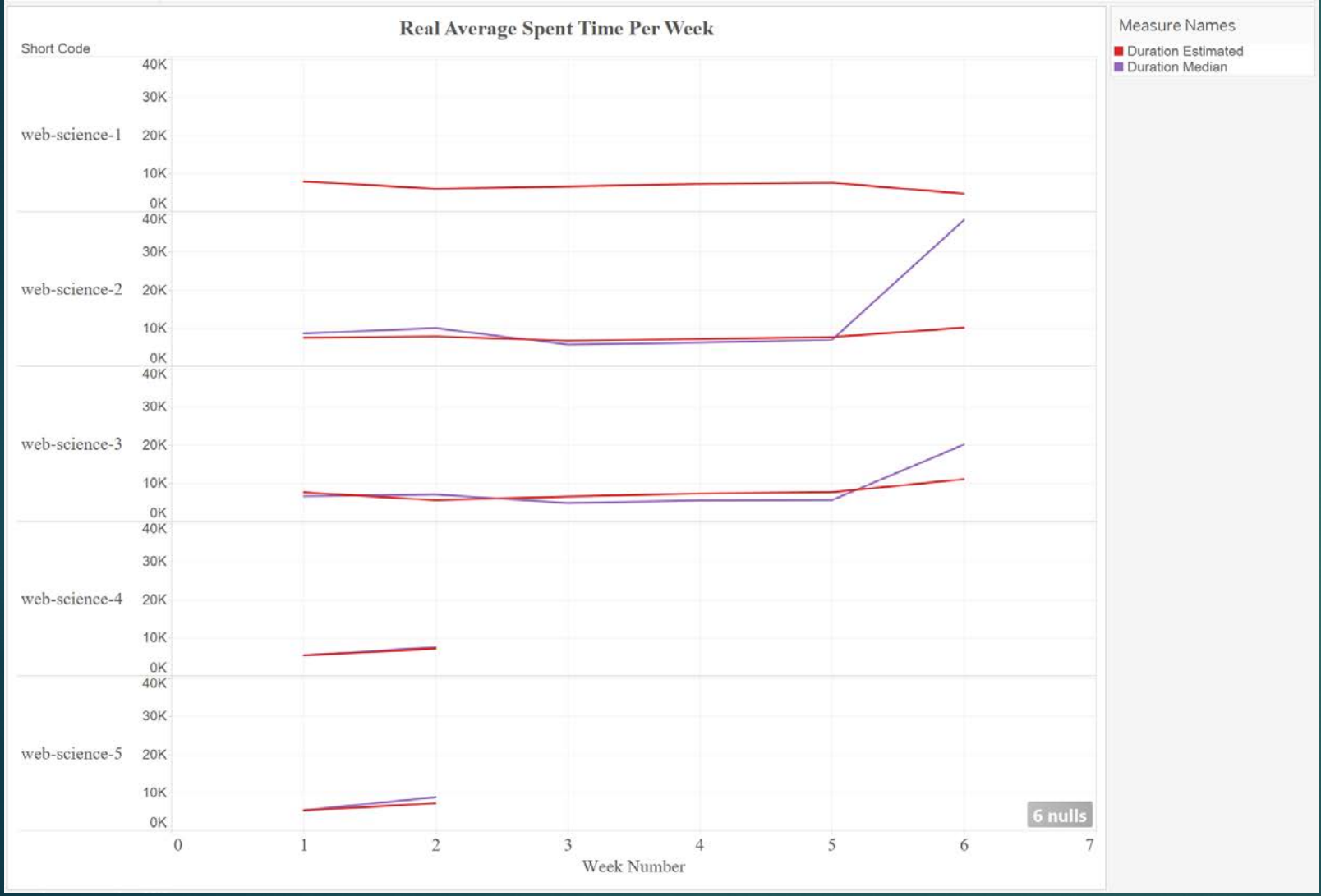
Learning Profiles – Duration Times



Courses

- **Research Project**
- Most times are over-estimated
- At design time, durations are not even, having a 2x difference between the peak (week 2) and lower point (week 4)
- Real duration variation may discourage the participant engagement with differences up to 10x among weeks. Roller Coaster pattern.

Learning Profiles – Duration Times



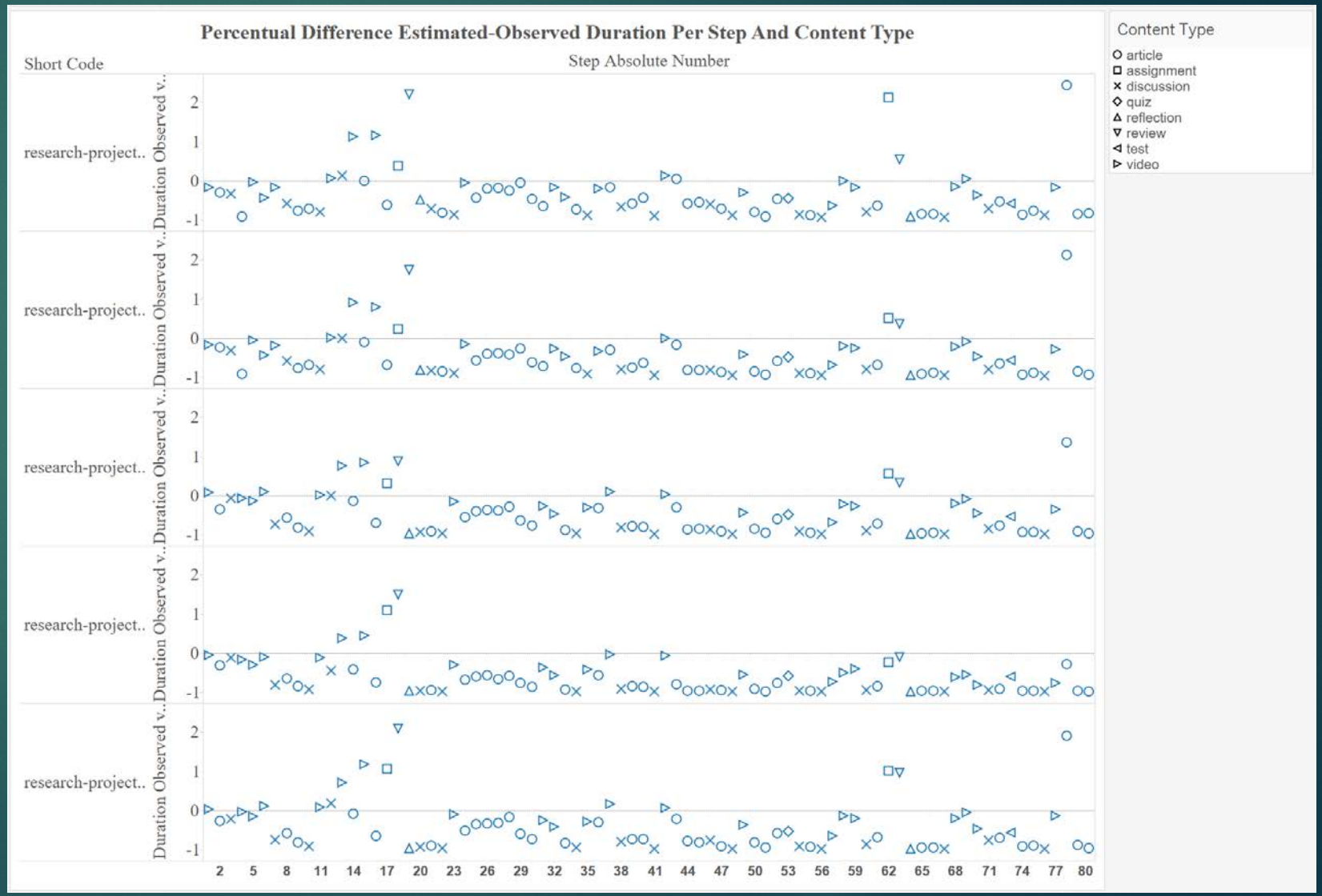
Courses

- **Web Science**
- Uses the same methodology as Research Project
- Times are more accurately estimated and observed during all course runs
- The last week difference is due to an underestimation of the assignment
- Overall the course seem to be better balanced in duration times

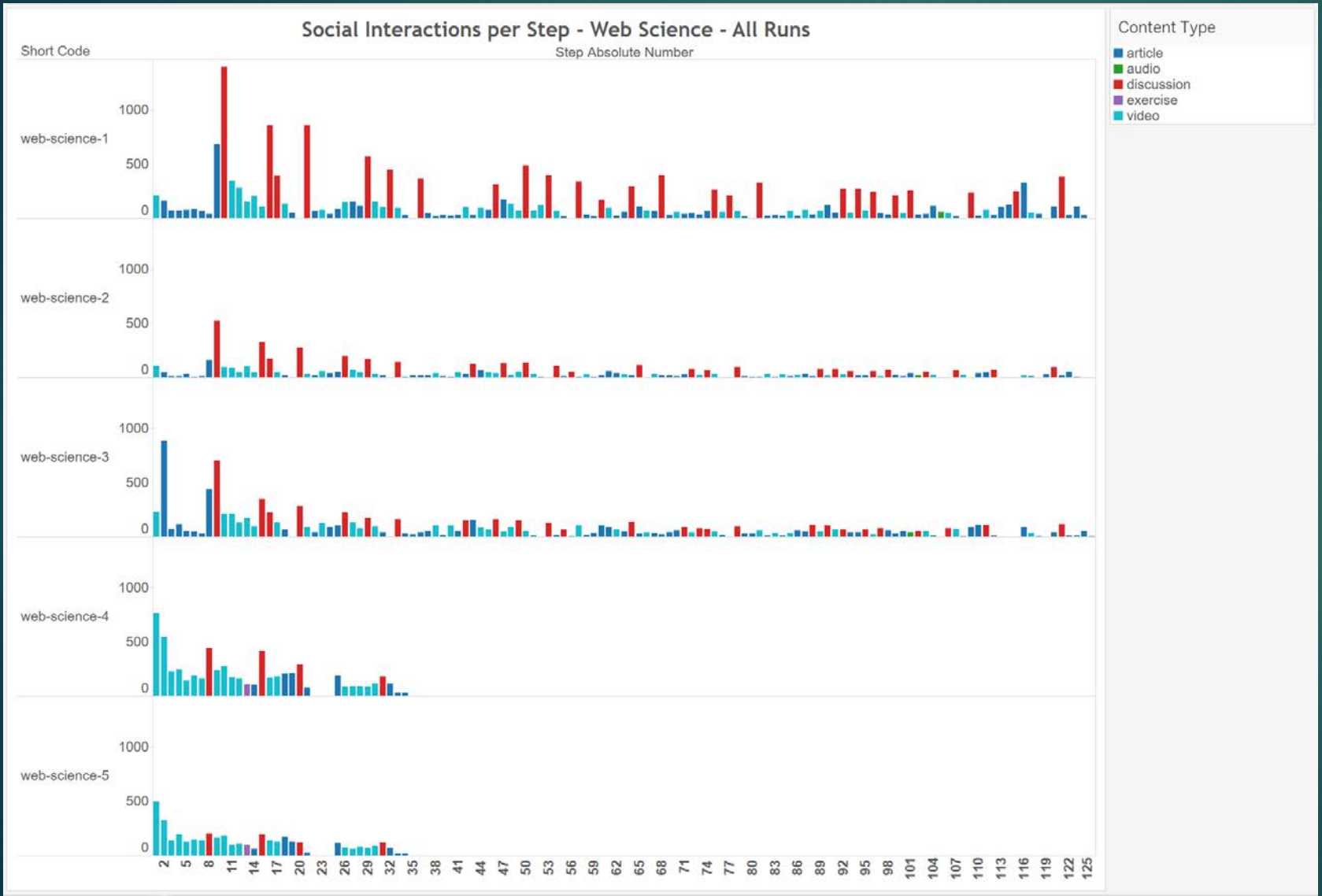
Learning Profiles – Duration Times

Courses

- In some cases, real vs estimated times vary significantly
- No particular content type was consistently found with higher rates of inaccuracy



Learning Profiles – Social Interactions



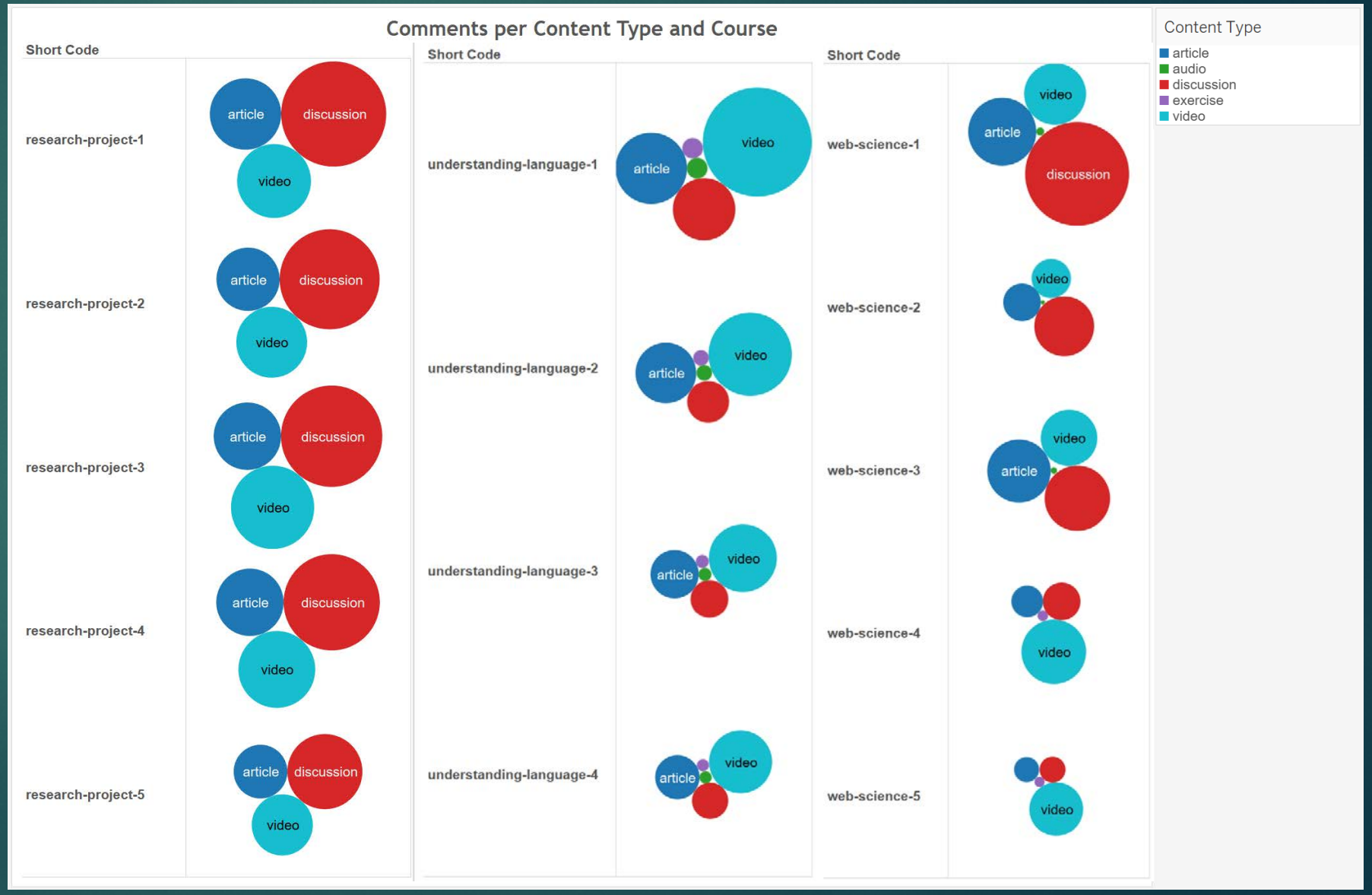
Behaviors

- **Web Science**
- Course design changed between runs 3 and 4 from 6 to 2 weeks
- Design change affected social interactions, specially for the Discussions content type (In red)
- Even if there were no important changes between runs 1, 2 and 3, run 1 clearly shows a higher rate of interactions. Some other factors may also impact social interactivity.

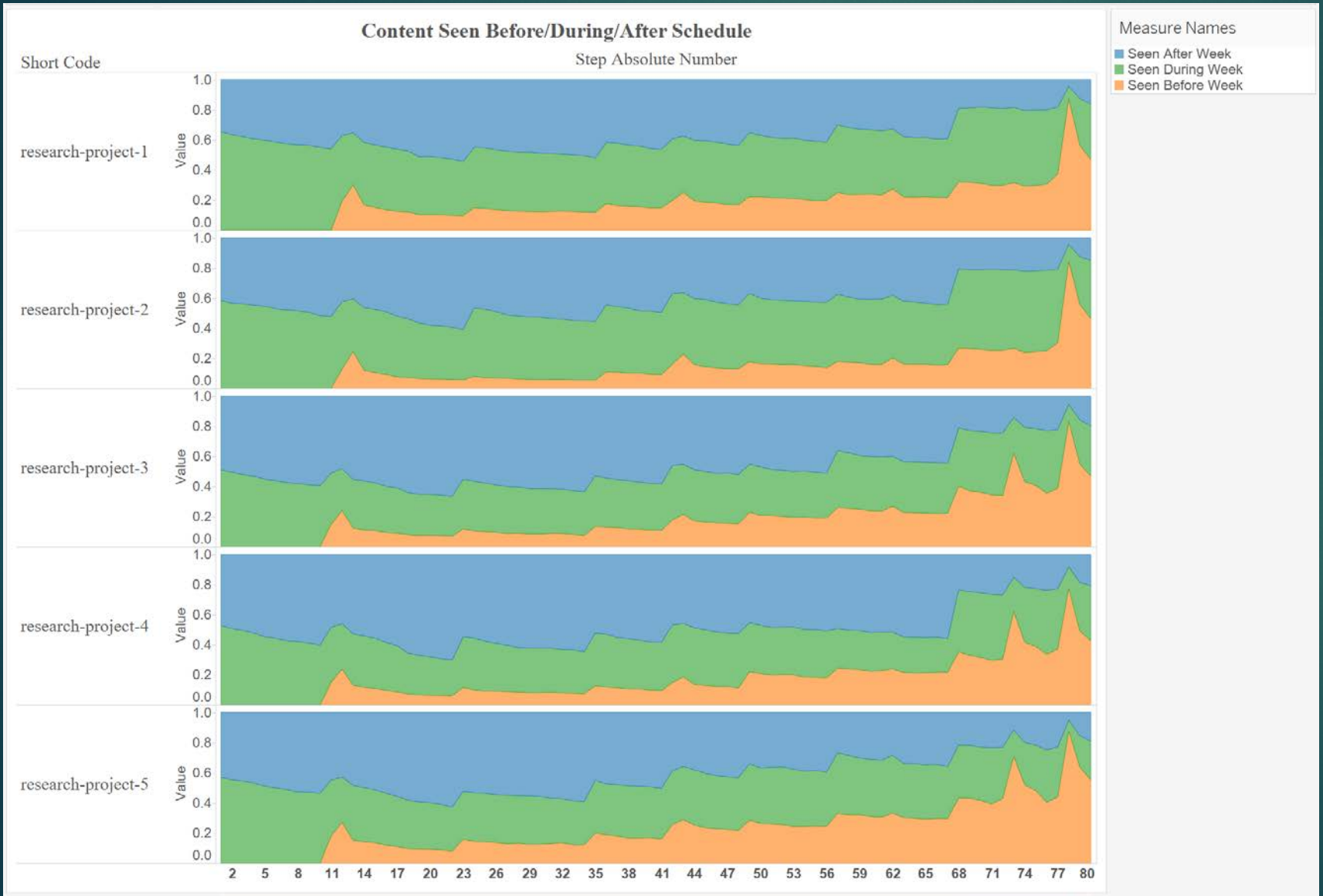
Learning Profiles – Social Interactions

Behaviors

- **All Courses**
- In all courses, the most socially active content types are discussion, video and article
- Course design seem to affect how effective is the content type “Discussion”



Learning Profiles – Content Consumption

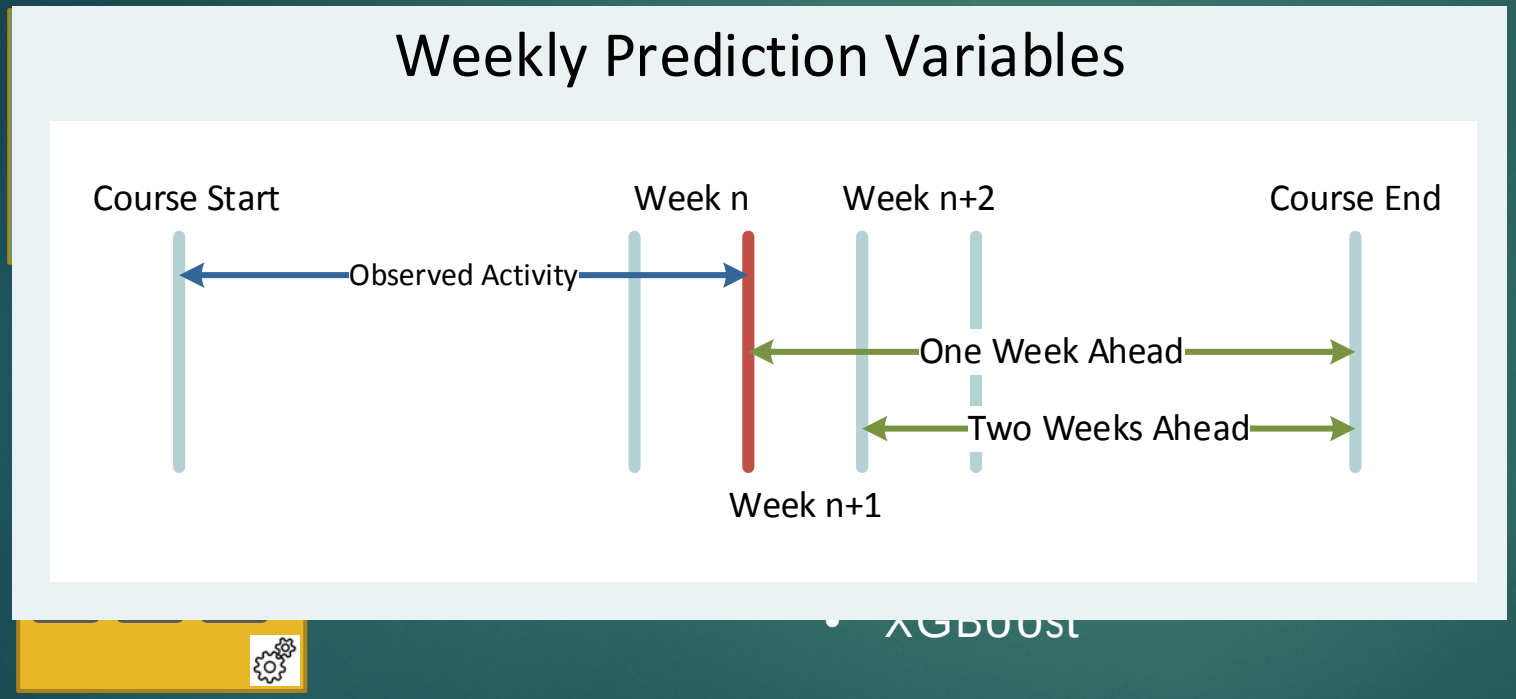


Behaviors

- Content consumed as scheduled is roughly 40% to 50% and almost constant after the week 1
- After week 2, content increasingly is consumed ahead of schedule and reducing almost the same proportion to delayed consumed content
- The reduced amount of participants at the later weeks of the course seem to be more proactive

Prediction – Dropout

Course Facts

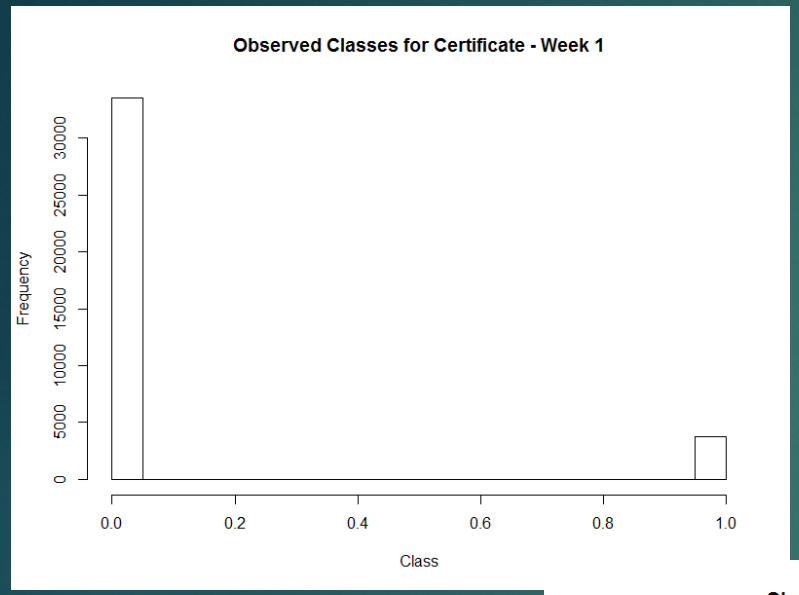


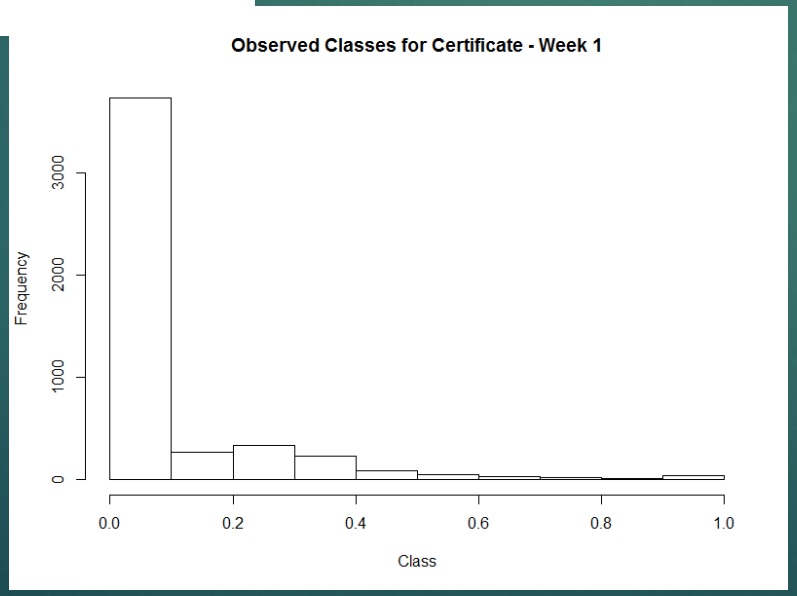
- XGBoost

To Predict

- Certificate: 50% steps completed + Assignments
- One Week Ahead: Any activity from the end of a selected week until the course end
- Two Weeks Ahead: Any activity from the end of the selected week + 1 until the course end

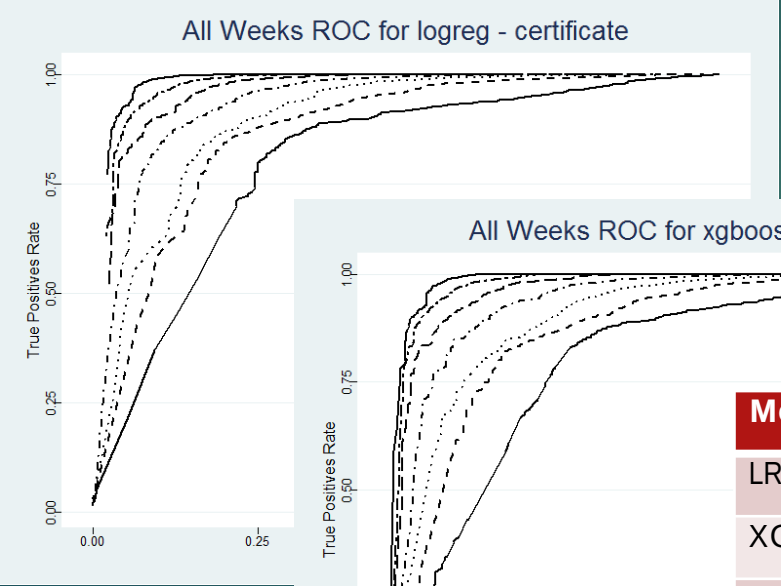
Prediction – Dropout

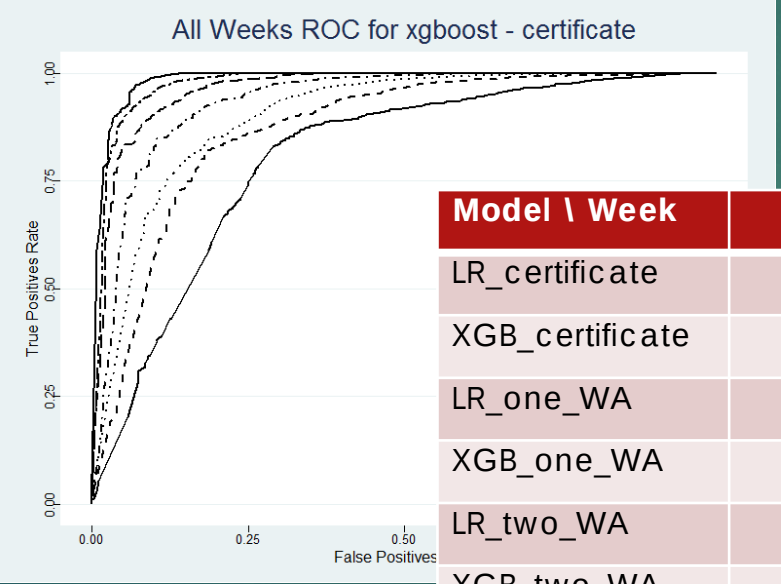




- Highly unbalanced classes
- Focus set on Specificity
- Use AUC to evaluate the models

Prediction – Dropout





Model \ Week	1	2	3	4	5	6	7
LR_certificate	0.8135	0.8748	0.8997	0.9373	0.9620	0.9740	0.9844
XGB_certificate	0.8052	0.8709	0.9000	0.9354	0.9638	0.9757	0.9851
LR_one_WA	0.6998	0.7247	0.7478	0.7639	0.7803	0.7944	0.8312
XGB_one_WA	0.7080	0.7304	0.7624	0.7741	0.7849	0.8025	0.8296
LR_two_WA	0.6819	0.7162	0.7438	0.7655	0.7844	0.8145	0.9844
XGB_two_WA	0.6951	0.7218	0.7508	0.7673	0.7874	0.8202	0.9803

Recommendations

Recommendations (1)

- **Content material times** require more attention in some courses to make them more balanced
- FutureLearn may **show weekly effort estimations in the interface** setting the right expectations
- **Higher granularity** is required in step-activity data to obtain more accurate spent times and navigation sequences
- **The interface may capture more user events** to get additional understanding of user activity
- **Demographics** are highly desirable to identify cultural, educational, language, age or other factor related patterns
- **Location** can be calculated at city/region level for each step-activity from the IP address. This is key to add more cultural context and calculate the time zone.
- The **device type** is one of the most important missing fields in step-activity. It helps understanding how limited users are to use all features and for course design.

Recommendations (2)

- Social interactions seem to depend mainly on **how the course is designed**. Identifying factors (with more data) and creating guidelines is important.
- The user interface is confusing sometimes with the **comments option**, depending on how wide is the browser window.
- In **social steps** it would be helpful to highlight the relevant controls or showing floating comments to encourage more participation
- With an improved version of the prediction model created within this project, implement a **proactive "user leave" identification feature** in the platform, so partners can target communications to reduce the dropout rates.

Future Work & Resources

Future Work

- Analytics functions
  - Add new models or improve the existing ones
  - Allow incremental updates for daily deltas
  - Associate or create specific visualization tool with pre-built reports
- Software Development
  - Wrap the solution as an R Package
  - Support different storage types
  - Increase scalability by supporting multi-processing frameworks

Resources

GitHub Repository
http://github.com/miballeuk/FutureLearnAnalytics