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

Courses

Participants

Behaviors

- develop your academic reading skills by practicing scanning a text for specific information and skimming a text to get the gist

- practice deconstructing and understanding an academic argument when reading in order to create an argument in your own writing

- join the discussion on how you can become proficient at note taking

© University of Southampton 2015
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
- Reviews
- Assignments
- Questions

Courses  Participants  Behaviors
<table>
<thead>
<tr>
<th>short_code</th>
<th>run_number</th>
<th>short_name</th>
<th>full_name</th>
<th>start_date</th>
<th>end_date</th>
<th>institution</th>
<th>department</th>
</tr>
</thead>
<tbody>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>Developing your research project</td>
<td>2014-07-07 00:00:00 UTC</td>
<td>2014-09-01 00:00:00 UTC</td>
<td>UoS</td>
<td>Student Recruitment</td>
</tr>
<tr>
<td>research-project-2</td>
<td>2</td>
<td>research-project</td>
<td>Developing your research project</td>
<td>2014-09-15 00:00:00 UTC</td>
<td>2014-11-10 00:00:00 UTC</td>
<td>UoS</td>
<td>Student Recruitment</td>
</tr>
<tr>
<td>research-project-3</td>
<td>3</td>
<td>research-project</td>
<td>Developing your research project</td>
<td>2015-06-22 00:00:00 UTC</td>
<td>2015-08-17 00:00:00 UTC</td>
<td>UoS</td>
<td>Student Recruitment</td>
</tr>
<tr>
<td>research-project-4</td>
<td>4</td>
<td>research-project</td>
<td>Developing your research project</td>
<td>2015-09-14 00:00:00 UTC</td>
<td>2015-11-09 00:00:00 UTC</td>
<td>UoS</td>
<td>Student Recruitment</td>
</tr>
<tr>
<td>research-project-5</td>
<td>5</td>
<td>research-project</td>
<td>Developing your research project</td>
<td>2016-03-07 00:00:00 UTC</td>
<td>2016-05-02 01:00:00 UTC</td>
<td>UoS</td>
<td>Student Recruitment</td>
</tr>
<tr>
<td>web-science-1</td>
<td>1</td>
<td>web-science</td>
<td>Web Science: How the Web is changing the world</td>
<td>2013-11-11 00:00:00 UTC</td>
<td>2013-12-23 00:00:00 UTC</td>
<td>UoS</td>
<td>ECS</td>
</tr>
<tr>
<td>web-science-2</td>
<td>2</td>
<td>web-science</td>
<td>Web Science: How the Web is changing the world</td>
<td>2014-02-10 00:00:00 UTC</td>
<td>2014-03-24 00:00:00 UTC</td>
<td>UoS</td>
<td>ECS</td>
</tr>
<tr>
<td>web-science-3</td>
<td>3</td>
<td>web-science</td>
<td>Web Science: How the Web is changing the world</td>
<td>2014-10-06 00:00:00 UTC</td>
<td>2014-11-17 00:00:00 UTC</td>
<td>UoS</td>
<td>ECS</td>
</tr>
<tr>
<td>web-science-4</td>
<td>4</td>
<td>web-science</td>
<td>Web Science: How the Web is changing the world</td>
<td>2015-11-30 00:00:00 UTC</td>
<td>2015-12-14 00:00:00 UTC</td>
<td>UoS</td>
<td>ECS</td>
</tr>
<tr>
<td>web-science-5</td>
<td>5</td>
<td>web-science</td>
<td>Web Science: How the Web is changing the world</td>
<td>2016-06-27 00:00:00 UTC</td>
<td>2016-07-11 00:00:00 UTC</td>
<td>UoS</td>
<td>ECS</td>
</tr>
<tr>
<td>understanding-language-1</td>
<td>1</td>
<td>understanding-language</td>
<td>Understanding Language: Learning and Teaching</td>
<td>2014-11-17 00:00:00 UTC</td>
<td>2014-12-15 00:00:00 UTC</td>
<td>UoS</td>
<td>Languages</td>
</tr>
<tr>
<td>understanding-language-2</td>
<td>2</td>
<td>understanding-language</td>
<td>Understanding Language: Learning and Teaching</td>
<td>2015-04-20 00:00:00 UTC</td>
<td>2015-05-18 00:00:00 UTC</td>
<td>UoS</td>
<td>Languages</td>
</tr>
<tr>
<td>understanding-language-3</td>
<td>3</td>
<td>understanding-language</td>
<td>Understanding Language: Learning and Teaching</td>
<td>2015-10-19 00:00:00 UTC</td>
<td>2015-11-16 00:00:00 UTC</td>
<td>UoS</td>
<td>Languages</td>
</tr>
<tr>
<td>understanding-language-4</td>
<td>4</td>
<td>understanding-language</td>
<td>Understanding Language: Learning and Teaching</td>
<td>2016-04-04 00:00:00 UTC</td>
<td>2016-05-02 00:00:00 UTC</td>
<td>UoS</td>
<td>Languages</td>
</tr>
<tr>
<td>short_code</td>
<td>run_number</td>
<td>short_name</td>
<td>week_number</td>
<td>step_number</td>
<td>content_type</td>
<td>duration_estimated</td>
<td>week_start_date</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>-----------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------------</td>
<td>--------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>1</td>
<td>video</td>
<td>240 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>2</td>
<td>article</td>
<td>180 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>3</td>
<td>discussion</td>
<td>300 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>4</td>
<td>article</td>
<td>420 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>5</td>
<td>video</td>
<td>240 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>6</td>
<td>video</td>
<td>240 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>7</td>
<td>video</td>
<td>240 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>8</td>
<td>discussion</td>
<td>300 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>9</td>
<td>article</td>
<td>600 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>10</td>
<td>article</td>
<td>180 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>11</td>
<td>discussion</td>
<td>300 2014-07-07 00:00:00 UTC</td>
<td>2014-07-14 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>1</td>
<td>12</td>
<td>article</td>
<td>240 2014-07-14 00:00:00 UTC</td>
<td>2014-07-21 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>2</td>
<td>1</td>
<td>video</td>
<td>240 2014-07-14 00:00:00 UTC</td>
<td>2014-07-21 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>2</td>
<td>2</td>
<td>discussion</td>
<td>300 2014-07-14 00:00:00 UTC</td>
<td>2014-07-21 00:00:00 UTC</td>
</tr>
<tr>
<td>research-project-1</td>
<td>1</td>
<td>research-project</td>
<td>2</td>
<td>3</td>
<td>video</td>
<td>60 2014-07-14 00:00:00 UTC</td>
<td>2014-07-21 00:00:00 UTC</td>
</tr>
</tbody>
</table>
Data Processing Pipeline

Facts!
The Analysis
• 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
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

- **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”
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

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

Weekly Prediction Variables

- Course Facts:
  - Observed Activity
  - One Week Ahead
  - Two Weeks Ahead

- Models:
  - LogRegression
  - XGBoost
Prediction – Dropout

- Highly unbalanced classes
- Focus set on Specificity
- Use AUC to evaluate the models
## Prediction – Dropout

### ROC Curves

- **All Weeks ROC for logreg - certificate**
- **All Weeks ROC for xgboost - certificate**

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