A Process for Predicting MOOC Attrition

Mike Sharkey
President
Blue Canary
Chandler, AZ USA
mike@bluecanarydata.com

Robert Sanders
Sr. Software Engineer
Clairvoyant, LLC
Chandler, AZ USA
robert.sanders@clairvoyantsoft.com

Abstract

The goal of this shared task was to predict attrition in a MOOC through use of the data and logs generated by the course. Our approach to the task reinforces the idea that the process of gathering and structuring the data is more important (and more time consuming) than the predictive model itself. The result of the analysis was that a subset of 15 different data features did a sufficiently good job at predicting whether or not a student would exhibit any activity in the following week.

1 Introduction

Blue Canary is a higher education analytics company located in Chandler, Arizona USA. The company has extensive experience in dealing with academic course/enrollment/retention data and is proud to collaborate with other researchers on the EMNLP 2014 shared task. The goal of the task is to use data from one MOOC, create a model to predict course attrition, and then apply that model to five other MOOCs in order to observe the efficacy of the model across courses. The goal of this paper is to document the process that Blue Canary went through in order to generate the model.

2 Understanding the Problem

In order to successfully complete a task such as this, the team needed the right context to the problem. The context for this particular challenge (using MOOC data to predict attrition) was very familiar to the Blue Canary team. First, the team has developed retention-oriented predictive models for a number of institutions in the past. This experience was vital. Second, the team has worked with data at scale. The MOOC course had 20,000 enrolled students with a log file that generated 1.6 million rows of data. The Blue Canary team has experience working with a large online university that had over 300,000 students generating millions of rows of data on a daily basis. Lastly, all of the team members have participated in at least one MOOC, so the processes and interactions associated with such a course are known.

The combination of all of these factors gave the Blue Canary team the necessary context to tackle the attrition problem from the ground up.

3 Approach to the Problem

As with other such data initiatives, the process is a stepwise iterative one. Each step and iteration provides more insight, allowing the team to refine the prediction.

3.1 Step 1: Feature Extraction

Feature extraction is the process of defining the independent variables (or inputs) for the predictive model. This is arguably the most important step in the process of developing a predictive model. It requires a deep understanding of the source data from a technical side as well as a contextual understanding of how the data relate to the front-end user experience.

Blue Canary used two techniques for feature extraction. The first was experience. Having looked at course activity data and developed predictive models for other courses, we knew the kinds of features that would likely have an impact on the prediction. This experience gave us simplistic features like "number of videos watched" and "total minutes spent in class" to more nuanced features like "attempted quiz without referring to other materials".

The second technique was using visualizations to explore data relationships. The team used the Tableau visualization tool to ingest course activity data and map it across users & weeks. Looking at these relationships visually helped to determine if we should include the features in the modeling or not.

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 50–54,
October 25-29, 2014, Doha, Qatar.

3.2 Step 2: Define Outcome/Prediction

Once the list of features have been developed, next step is to define exactly what it is we are predicting. At a high level, it sounds easy – will the student retain in the class? From a data perspective, though, we need to define what it means to retain. Does it mean that the student submitted the assignment for the week? Watched a video? Simply logged in? Zeroing in on a reliable definition of retention is a part of the process.

3.3 Step 3: Run the Predictive Model

With the input and output data in place, the team needs to run a model to derive a prediction. Blue Canary has consistently used machine learning techniques (as opposed to statistical modeling). As Bogard (2011) alludes to in a blog post comparing the two approaches, Blue Canary's technical expertise combined with an unknown underlying relationship make machine learning our preferred method of analysis. For this analysis, Blue Canary implemented a random forest method using the SciKit python toolset (http://scikit-learn.org/).

3.4 Step 4: Observe/Validate/Iterate

The last step in the process is to observe the outcomes of the modeling, validate the results (both quantitatively and qualitatively) and iterate to improve. When looking at the modeling results, we focused on accuracy. More specifically, we focused on the true positive rate (recall) and the true negative rate individually. The combination of these components equal the accuracy of the model, but we thought it was important to look at both since the application of any such solution would involve treatments for both parties.

Value	Definition
True Positive	# predicted to retain / # actually retained
True Negative	# predicted to attrite / # actual attrition
Accuracy	(True positive + True negative) / population

Table 1: Definition of model accuracy values

3.5 Acknowledging Prior Research

It should be noted that Blue Canary has stood on the shoulders of others who have tackled similar problems in the past. Our choice for analytical methods and features has been inspired by earlier predictive projects like Purdue's Course Signals (Arnold and Pistilli, 2012) and research done at American Public University (Boston et. al., 2011). We also referenced contemporary MOOC research that explored the descriptive (Breslow et. al., 2013), predictive (Taylor et. al., 2014), and social (Rosé et. al., 2014) contributors to attrition.

4 Predicting Attrition for PSY-001

The course in question was from a 2013 Georgia Tech/Coursera MOOC called "Introduction to Psychology as a Science". Blue Canary executed seven iterative steps as explained in the previous section. At the end we came up with a model that used 15 features to predict retention and attrition at an 88% accuracy rate.

4.1 Iteration 1: Feature Extraction

The first iteration didn't result in any prediction. The goal was to explore the data and extract an initial set of features for processing. We also created our training, testing, and hold back data using a 70/15/15 split. Table 2 lists the features we initially extracted from the activity data.

- id
- user_id
- username
- week_id
- week_num
- week_start_date
- week_end_date
- session_count
- url_wiki_edit_count
- url_wiki_view_count
- url_quiz_count
- url_lecture_count
- url_forum_count
- is_english
- ip_count
- most_common_browser
- most_common_browser_date
- browser_count
- unique_quizzes_attempted
- total_quiz_attempts
- average_attempts_per_quiz
- videos_accessed_count
- average_video_per_session
- did_peer_review
- actually_attended

Table 2: Initial list of features

These features were very basic. We didn't spend much time on more advanced features. The goal of this first was simply to lay the foundation for our data analysis pipeline.

4.2 Iteration 2: Test Analytical API's

With a bulk of the features in place, our next goal was to connect the machine learning toolset to the pipeline. We used Weka (http://www.cs.waikato.ac.nz/ml/weka/) since the team had some experience with the tool. Since our approach was to construct the pipeline as a smooth-running application, we utilized the Weka API's to feed data in and get results out.

Unfortunately, we ran into technical problems with the API's and got out of memory exception errors. We were unable to troubleshoot and decided to move on to another toolset. In addition, though, we added more features, mainly from parsing the URL strings in the access log files (Table 3).

- event_count
- total_minutes_spent
- url_quiz_submits_count
- url_quiz_actual_submits_count
- url_quiz_percent_of_actual_submits
- url_quiz_attempt_in_more_than_one_session
- url_quiz_retry
- url_quiz_attempt_but_no_submit
- url_quiz_submit_no_help
- url_human_grading_count
- url_forum_search_count
- url_class_preferences_count
- url_signature_count

Table 3: URL features added

4.3 Iteration 3: Too Good to be True

We switched to SciKit as our analytical tool of choice, but we still used the Random Forest method. We ran our first analysis and got the corresponding accuracy rates. As explained in section 3.4, we produce accuracy rates for 'False' (correctly predicting that the student won't attend next week), 'True' (correctly predicting that the student will attend next week) and 'Average' (accuracy – the weighted average of False and True). The results for our first run were as follows:

Measure	Rate
False	99%
True	87%
Accuracy	96%

The team was skeptical about such high accuracy rates, especially given that it was our first run. We suspected that there was some sort of leakage – information about the prediction field may have leaked into one of the features. That suspicion was confirmed when we dug deeper into the model.

The predominant feature was "is_english". We looked at the user agent data in the activity logs and parsed the language parameter to determine if the web browser language was set to English or not. It turns out that when there was no activity for the week, we populated this field with null values. Since the majority of the students had English as their language, the model was seeing "is_english" = TRUE when there was activity and "is_english" = FALSE when there wasn't activity. This was a great example of the kinds of errors one finds early on in the analysis.

4.4 Iteration 4: First Real Model

For the next iteration, we fixed the "is_english" field and ran the model again. This run was our first valid predictive model for the dataset and the results were:

Measure	Rate
False	92%
True	55%
Accuracy	89%

Note that we are doing a very good job at predicting students who won't attend next week. This is due to the fact that there are a large number of students don't attend. We estimated that about 20,000 students signed up for the class, 11,000 of them showed any activity at all, and less than 3,000 completed the course.

4.5 Iteration 5: Defining the Outcome

For experimentation purposes, we wanted to see if changing the definition of "attending" would have any effect on the modeling. Our original definition of attending was that there were ANY user actions in the data (viewing a page, posting a discussion item, taking a quiz, etc.). We decided to add variations to that definition such as "viewing at least one lecture", "submitting at least one quiz", or "will never attend again" (as opposed to

just not attending next week). The table below is a sampling of some of the results we generated:

Measure	Out_i	Out_a	Out_b	Out_c
False	92%	94%	97%	87%
True	55%	45%	47%	90%
Accuracy	89%	91%	95%	89%

This exercise showed some interesting results. Specifically, we saw how we would improve our ability to predict students who wouldn't attend (False) but decrease the True accuracy. We did see significant improvement in the case where the outcome was "will never attend again". However, we decided to stay with our base definition of attendance as "no activity in the following week". Validating these alternate definitions of attendance is a task that would be worthwhile for additional research.

4.6 Iteration 6: Team Collaboration

Blue Canary prides itself on collaboration not only amongst researchers in the learning analytics field, but also collaboration inside of our own company. We made sure to share information about this shared task with others in the company, and that collaboration allowed us to positively expand our feature set. One employee had come across MOOC research that had found good predictive results when using an aggregate engagement/activity score (Poellhuber, 2014). We decided to utilize a similar feature where the number of sessions, pages, days, and hours of activity in a given week were combined into an engagement score.

4.7 Iteration 7: Winnowing the Field

As a final step, we wanted to reduce the number of features used in the modeling process so as to improve cycle times. We knew that the majority of the fields had little to no predictive value, so we ran models where we just used the top 10, 15, or 20 features. In the end, all permutations gave similar accuracy scores and we decided to use the top 15 features. Those features resulted in accuracy rates of:

Measure	Rate
False	92%
True	54%
Accuracy	88%

The accuracy rates are similar to the rates we had been getting in the past two iterations of the modeling. This led us to conclude that we were at the point of diminishing returns and we decided to finalize the model with the 15 features and their corresponding importance level as illustrated in Table 4 (below).

Feature	Import.
total_minutes_spent_previous_wk	0.336
initial_activity_score_previous_wk	0.072
final_activity_score_previous_wk	0.071
final_activity_score_up_to_wk	0.070
event_count_up_to_wk	0.068
most_common_browser_count_up_to_wk	0.059
initial_activity_score_up_to_wk	0.049
url_wiki_view_count_up_to_wk	0.041
session_count_up_to_wk	0.038
url_quiz_count_up_to_wk	0.037
total_minutes_spent_up_to_wk	0.037
url_lecture_count_up_to_wk	0.037
browser_count_up_to_wk	0.031
ip_count_up_to_wk	0.031
session_count_previous_wk	0.023

Table 4: Features and Importance

5 Conclusions

The overarching conclusion from this research can be summarized in two points:

1. Machine learning models can do an above average job at predicting retention/attrition in MOOC's
2. The predictive factors are not surprising – they are variants of measures of the student's engagement and activity in the course

5.1 Features

Looking at the features in Table 4, one can see that almost all of the important features are measures of activity. Minutes, events, views and even the aggregated activity feature are all measuring similar characteristics. The takeaway here is that there shouldn't be an expectation of some unique marker that predicts retention. There's no secret in the secret sauce.

6 Acknowledgements

The authors would like to thank Mohammed Ansari, Andy Allen, Satish Divakarla, David Morgan, and the entire Blue Canary and Clairvoyant team for their support in this shared task.

References

Matt Bogard. (2011, January 29) Culture War: Classical Statistics vs. Machine Learning. Retrieved from http://econometricsense.blogspot.com/2011/01/classical-statistics-vs-machine.html

Arnold, K. E., & Pistilli, M. D. (2012, April). Course Signals at Purdue: Using learning analytics to increase student success. In Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (pp. 267-270). ACM.

Boston, W. E., Ice, P., & Gibson, A. M. (2011). Comprehensive assessment of student retention in online learning environments. Online Journal of Distance Learning Administration, 14(4).

Breslow, L., Pritchard, D. E., DeBoer, J., Stump, G. S., Ho, A. D., & Seaton, D. T. (2013). Studying learning in the worldwide classroom: Research into edX's first MOOC. Research & Practice in Assessment, 8, 13-25.

Taylor, C., Veeramachaneni, K., & O'Reilly, U. M. (2014). Likely to stop? Predicting Stopout in Massive Open Online Courses. arXiv preprint arXiv:1408.3382.

Rosé, C. P., Carlson, R., Yang, D., Wen, M., Resnick, L., Goldman, P., & Sherer, J. (2014, March). Social factors that contribute to attrition in moocs. In Proceedings of the first ACM conference on Learning@ scale conference (pp. 197-198). ACM.

Poellhuber, B., Roy, N., Bouchoucha, I., Anderson, T. (2014, April). The Relationship Between the Motivational Profiles, Engagement Profiles and Persistence of MOOC Participants. Retrieved from http://www.moocresearch.com/wp-content/uploads/2014/06/MOOC-Research-InitiativePoelhuber9187v4a.pdf, September 1, 2014.