Modeling MOOC Dropouts

Pratul Ramkumar
Computer Science
PES University
Bangalore,Karnataka
pratul.ramkumar@gmail.com

Siddhant Gupta
Computer Science
PES University
Bangalore,Karnataka
dpsrkp.sid@gmail.com

Yash Agarwal
Computer Science
PES University
Bangalore,Karnataka
yagarwal095@gmail.com

I. Abstract

In this project, we are trying to model MOOC dropouts. Several previous work done to address this issue, has dealt with the forum interactions, and discussion module to get results. However we avoid using this approach as only 5% to 10%(Rose and Siemens, 2014[1]) of the students enrolled in an MOOC course use the forum, and instead look at clickstream or click based interaction with the platform. We generate several features, based on the click stream activity and user interaction with various course modules - such as page navigations, problems accessed, wiki access. Also number of server side and browsers side requests fired by each user. We then have a classification problem at hand, and use - Naive Bayes, Decision Trees, KNN and Logistic regression classifiers. Our best model was obtained with the Logistic Regression of weekly activity counts.

II. Introduction

Massive Open Online Courses(MOOCs) have revolutionised the education industry, with millions of students already registered, and new, unique courses being added at a rapid rate, some suggest that MOOCs will bury higher education as students will opt supersized online courses over traditional classrooms. However, this is a distant vision, currently MOOCs face a critical issue - Dropouts.

In 2011, Andrew Ng adapted his Stanford Machine Learning class into a free online course for nearly 104,000 students. Ng hailed the numbers as a lifetime achievement. However, only 44 percent of the whooping number attempted the first assignment and only 13,000 students finished the class.[2]

With almost no cost of registration and a very high student to teacher ratio, completion rates for MOOCs are dismal, ranging between 7 to 10 percent[3]. It is hence a viable option to explore the dropout rates of these open courses, and possibly even uncover certain incentives, approaches to encourage students to complete the course.

III. Summary of literature survey report

In the literature survey report, we had discussed in detail work done to predict MOOC dropouts by gaining insights into forum posts and forum click stream data - Ramesh et al[5]. The paper considered posting patterns and number of upvotes, downvotes among forum interaction. The researchers also delved into linguistic features such as polarity and tone of posts on the forum to assess how the user was going about the course modules. These features were finally combined using Probabilistic Soft Logic(PSL) rules, to conclude if the user completed a course or not.

Yang et al[4], considered the number of threads and sub-threads that a user has started or has been a part of. The work including looking at the average number of words used on the forum, post density - number of posts in a week. The motivation behind this was to evaluate whether forum starters are survivors. Or is posting on the forum even related to finishing a course.

Finally, Sinha et al[6] tried to predict dropouts by analysing the clickstream on video modules, such as pausing, playing, forwarding, skipping video lectures.

IV. Problem Statement

Most of the work as discussed above is contextual and related to the forum. We believe that this approach is far to restrictive, as only 5% to 10% of the students enrolled in a MOOC use the forum (Rose and Siemens, 2014[1]). We hence look at count based features such as page access, page navigation, wiki access, problems accessed etc. MOOCs have challenged the conventional education system in every way, however, themselves face an embarrassing and currently inevitable issue of very high Dropout Rates. We hence delve into the matter, considering user activity data. It is important to predict dropouts. Our goal is be to understand shared practices in the user click data, and how can hidden patterns in user study behaviour be used to develop a healthier and conducive learning environment.

V. Data Description

The dataset used for this work has been sourced from the KDD Cup Challenge 2015 [9]. The data is entirely relation and event based, with no contextual information. The details of the initial dataset are as follows:

object/module data - *<course_id, module_id, category, children, start>* contains data for each course having several modules.

course_id = ID for a particular course

module_id = each course can have multiple modules, with a module ID

category = category of a module can be - chapter, course_info, about, sequential, vertical, discussion, outlink, static_tab, peergrading, course, combinedopenended, html, video, dictation , problem

children - module ID's for all children for a particular module

start - start date for a module

The object module contains several duplicate entries, that is the same set of <course_id, module_id, category> are repeated in the file. We hence processed the object file to obtain the 'objectunique' file, which contains a single entry for the <course_id, module_id, category> triplet.

log data - *<enrollment_id, time, source, event, object>* contains data for an event on the page of a particular module.

enrollment_id = enrollment ID of the student, the enrollment ID for the same student is different for different courses.

time = time of event

source = server, browser

event - access, discussion, navigate, page_close, problem, video, wiki

object - ID of the module

The log data contains several objects which have no listing in the object data file.

truth train data - *<enrollment_id, result>* whether the student completed the course or not.

enrollment_id = enrollment ID of the student.

result = 1 indicates that the student dropped out, and 0 indicates that the student completed the course.

VI. Workflow of our approach

Figure 1 shows our approach to Modeling dropouts in the KDD Cup dataset.

VII. Feature Extraction and Modelling a Predictive System

Our work was broadly divided into two parts. In Part 1, we specifically look into the type of clickstream activity, looking at the activities for the entire duration of the course.

In Part 2, we break down the user activity into weeks, for each week we look at the total activity count of the user, not looking at the specifics. In this section we discuss about the Extraction of Features, Visualisation, and Modelling in both the parts.

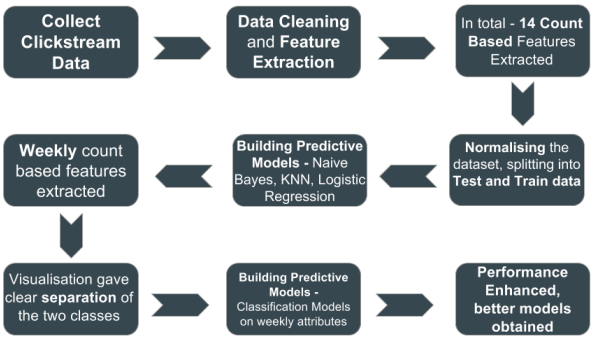

Fig. 1. Flowchart

Part 1

Feature Extraction

We obtained 14 features from the data sets listed above, after processing the data we obtained - *<enrollment_id,access,discussion,navigate,page_close, problem, video, wiki, server, browser, chapter, unknown, sequential, tot_time, session, result>*

From each entry of **enrollment ID** in the log data we obtained the category of the object listed in the log record. The only non-zero features obtained from this grouping were number of chapter accesses and number of **sequential** accesses, other modules in the log data, but unavailable in the module data, were categorised as **unknown**.

The log data was itself used to form the heart of the clickstream attributes, each enrollment ID was grouped to obtain, the number of times a user has **accessed** a page of any module, the number of **discussions** participated in, total page **navigations**, count of **page closes**, number of **problem**, **video** and **wiki** accessed. We also include the number of **server** and **browser** requests issued, by grouping the log data.

A **session** is the period of time when the user is active. We look at fifty minute time windows, from the start of a user's log record. Any click activity that the user performs is considered to be part of one session. If a user is inactive for fifty minutes, the session expires, and any further logs for the user are part of the next session. The **total time** is the total length of all the sessions put together. It needs to be understood that a session may be less than or more than fifty minutes, but a session expires when the gap between two consecutive click records for the user is greater than fifty minutes.

Finally the **result** column was added, indicating whether a user dropped out - indicated by a 1 or not - indicated by a 0.

All the above mentioned 14 features were then normalised, limiting them to a **range of 0 to 1**.

We hence obtained a cleaned data set, for analysis, containing 120542 entries and 16 columns. This was followed by the creation of the train and test data sets. We randomly sampled the created features data set into two parts, following the Pareto Principle[7], which is the 80-20 rule, with eighty percent records in the training data and the remaining twenty percent in the test data. The resulting **training** data contained **96434** rows, and the **testing** data contained **24108** rows.

Visualisation

We performed various visualizations techniques to get a brief idea of the data and the relation between various features. We started with constructing the correlation matrix and the heat map of all the extracted features.

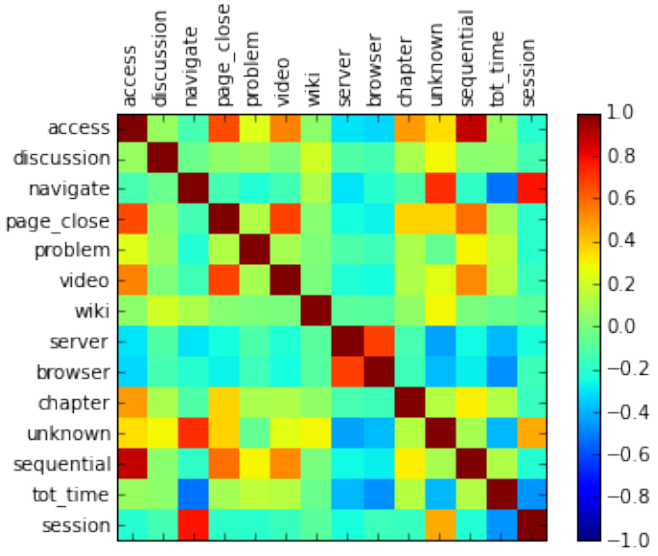

Fig. 2. Pairwise Correlation Heatmap

From Figure 2, we see that access and sequential are highly correlated, also session and navigate look to have a high correlation. However, later when forming the model, we establish that removal of either of the features for the development of the model does not give any improvement in performance. The attributes were hence left as it is.

Principal Component Analysis was performed, keeping only two of the most significant Eigenvectors. The 2-D features were then plotted, the results are as shown in Figure 3.

The points in Figure 3 indicate students, a red indicates that the student has dropped out, and blue indicates the user has successfully completed the course.

Roughly we see that the non dropouts are concentrated on the upper right side of the graph. Having so many (close to ninty thousand) data points in the dataset makes it hard to visualise the data efficiently. We hence see all the points clustered together, with no clear demarcation for any of the groups.

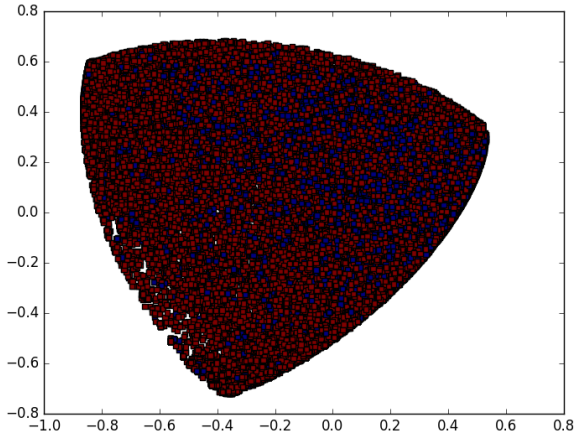

Fig. 3. Fourteen Attributes - PCA to Two Dimensions

Modeling Dropouts

After successfully creating a clean dataset, we had reduced our problem to a classification problem. That is we had to classify based on the fourteen extracted features whether the result column for a particular user would have a 0 - non dropout, or a 1 - dropout. We considered four classification models for the same.

1) We first considered the Naive Bayes classification approach. The model built on the train data yielded only 67.61% accuracy on the test data.
2) Next we built a Decision Tree classifier, the approach yielded an accuracy of 77.53%
3) We followed the decision tree classifier with a Logistic Regression based approach, as our final result was in a binary form. The Logistic Regression yielded an accuracy of 79.75%
4) Finally we used the K - Nearest Neighbor based approach, this resulted in a 80.56% efficient model.

Part 2

Feature Extraction

In this part we divided the users activity based on weeks, we go back to the original dataset to obtain a count of the the weekly activity for each user. <**enrollment_id, Week1, Week2, Week3, Week4, result**> For each **enrollment ID**, we obtain a count of the activity on the course website, irrespective of the module and type of interaction. The **result** column indicates whether the user is a dropout or not.

Visualisation

We plotted a graph for each user, depicting their weekly activity. Figure 4 shows the graph for user with enrollment ID - 1. User 1 was a non dropout, and had successfully completed the course.

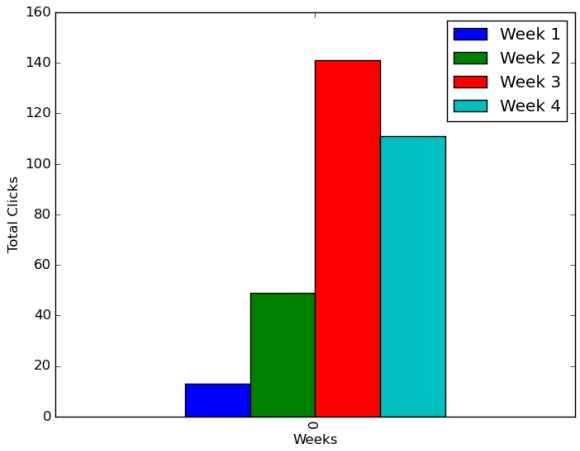

Fig. 4. User 1 Weekly Clicks

On picking a random enrollment ID - 1579, who turns out to be a dropout, the graph in Figure 5 indicates the dismal interaction the user has had with the course modules, the clicks peak in one week and fall greatly the next.

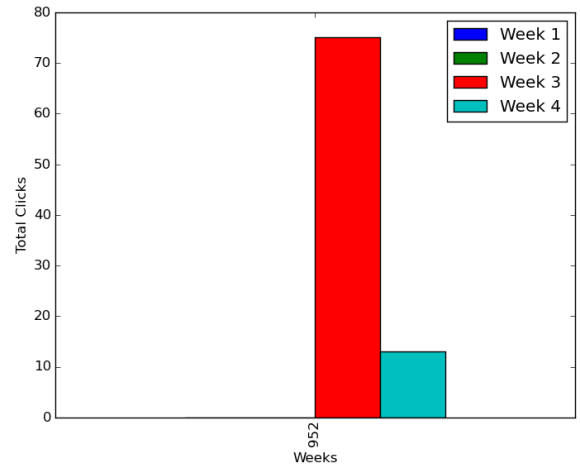

Fig. 5. User 1579 Weekly Clicks

Also, applying PCA to the data in the fourth dimension to reduce the attributes to three dimensions, we obtain some degree of separation in terms of dropouts and non dropouts, Figure 6, shows the dropouts in blue and non dropouts in red. Most of the non dropouts are concentrated towards the right at the tip of the graph.

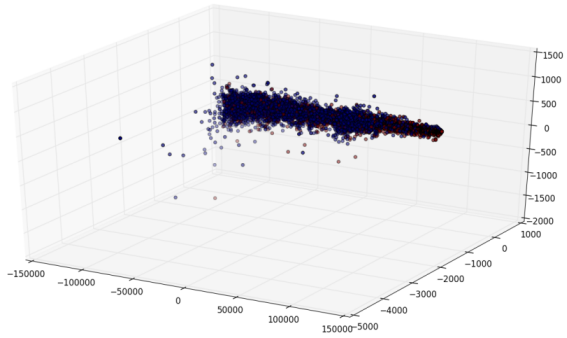

Fig. 6. Weekly Clicks - PCA to Three Dimensions

Modelling Dropouts

To model the data in this round, we divided the data into a train and test data set, with the same enrollment_id's in the train and test sets as the previous part. We then tried to model the result column, based on weekly activity counts. The same classifiers were used in this round as the previous. All the models, worked better in this Round, and significant improvements were obtained as compared to the previous round.

1) Naive Bayes classifier showed the most significant improvement, jumping to a whooping 84.60% accuracy from a mere 67% accurate model.
2) Decision Trees yielded 83.79% accuracy on the test data.
3) Logistic Regression correctly predicted dropouts to a 85.72%
4) And finally the KNN based classifier gave a 85.26% accurate model.

VIII. Experiments and Results

After building models to predict the results in each of the two rounds we tried to dig deeper into the results obtained. In Part 1 we now asked the question - Which of the 14 attributes does dropping out depend on the most? To answer this we found the coefficients from the Logistic Regression based classifier. Followed by finding the natural log of odds of success. From among the 14 click based attributes it was seen that **navigate** and **video** were ones the final result depended on most heavily. That is it is not enough to only watch videos. Other interactions are also required. For instances we observed that a student with chapter accesses is less likely to drop out than those who don't access the **chapter** modules. A student who does not access a chapter module, is 92% likely to dropout, but a student who access a chapter, is 61% likely to dropout.

Also finding the coefficients, and natural log success odds for Logistic Regression in Part 2. It was seen that Week 1, contributes the least towards dropping out. But Week 2 is the most significant contributor, closely followed by Week 4. It is hence a viable option to tap in at the second week and predict who is likely to dropout. The percentage change

in the number of clicks between the first and second week will indicate which user will probably not go on to finish the course. A small drop or an increase is likely to indicate that the user will finish the course. A huge slump will indicate that the student is losing interest or is finding it hard to understand the course modules.

Also for non dropouts specifically, the last weeks seems to be very busy. It can be seen from the following pie chart in Figure 7, that the highest number of average clicks for non dropouts are in the fourth week. The average clicks in numbers are - 43, 40, 46 and jumps to 66 in the last week. The Weekly activity peeks in Week 4 at 33.8% compared to the lower twenties in the initial weeks.

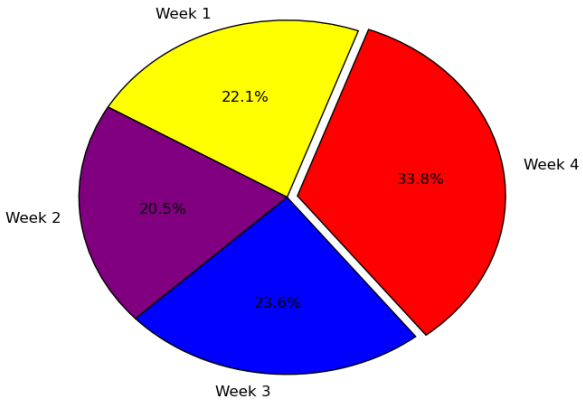

Fig. 7.

This goes to show that dedicated students buckle up as the course proceeds, or as the modules get tougher.

It is also important here to understand that although the average number of clicks fall in the second week, it is the percentage fall for every student which determines if they are likely to dropout in that or the future weeks.

IX. Conclusion

We here proposed a model to predict MOOC dropouts, based solely on click based features and counts. Several classification algorithms were used to decide if a given student, is a dropout or not. Including Naive Bayes, Decision Trees, KNN and Logistic Regression. We obtained good results based on the counts using a Logistic Regression based classifier, with about 85.26% accuracy. Also, we established that chapter access is crucial. A student who access chapters and other course material such as the course forum, is less likely to fall short on the fundamentals in the initial weeks. From the pie chart in the previous section we saw that Week 4 is a busy week for students who finish the course - non dropouts, dedicated students.

In the future we would look to combine both the Parts performed here, and generate a weekly count of the type of interaction, the weekly number of sessions for a user etc. and hope to improve the model further.

X. References

1. *Challenges and Opportunities of Dual-Layer MOOCs: Reflections from an edX Deployment Study* by **Carolyn Penstein Rosé, Oliver Ferschke, Gaurav Tomar, Diyi Yang, Iris Howley, and Vincent Aleven**

2. https://www.nas.org/articles/why_do_students_drop_out_of_moocs

3. https://en.wikipedia.org/w/index.php?title=Massive_open_online_course

4.*"Turn on, Tune in, Drop out": Anticipating student-dropouts in Massive Open Online Courses* by **Diyi Yang Tanmay Sinha David Adamson Carolyn Penstein Rose**

5. *Learning Latent Engagement Patterns of Students in Online Courses* by **Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, Lise Geetir**

6. *Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions* by **Tanmay Sinha, Patrick Jermann, Nan Li, Pierre Dillenbourg**

7. https://en.wikipedia.org/wiki/Pareto_principle

8. *ClaimEval: Integrated and Flexible Framework for Claim Evaluation Using Credibility of Sources* by **Mehdi Samadi, Partha Talukdar, Manuela Veloso, Manuel Blum**

9. https://kddcup2015.com

XI. Team Member contribution

1. **Pratul Ramkumar** - Data Cleaning, Building Predictive Models, Significant contribution in Part 2 of the Analysis, Documentation.
2. **Siddhant Gupta** - Visualisation, Data Cleaning, Significant contribution in Part 1 of the Analysis.
3. **Yash Agarwal** -Building Predictive Models, Visualisation, Significant contribution in Part 1 of the Analysis, Documentation.