Predicting student success based on interactions with Virtual Learning Environment

Vivek Doijode, Neha Singh

MS in Business Analytics, SAS and OSU Data Mining Certificate, Oklahoma State University

Introduction

Online learning can be called the millennial sister of classroom learning; tech savvy, always connected, and flexible. These features offer a convenient alternative to students with constraints and working professionals to learn on demand. According to National Center for Education Statistics, over 5 million students are currently enrolled in distance education courses. The growing trend and popularity of MOOCs (Massive Open Online Courses) and distance learning makes it an interesting area of research. We plan to work on OULA (Open University Learning Analytics) dataset. Learning analytics provides many insights on the learning pattern of students and on module assessments. These insights may be researched to enhance participants’ learning experience. In this paper, we predict students’ success in an online course using regression, clustering and classification methods. We have a mix of categorical and numeric inputs present in the OULA datasets that are in csv file formats and contain information for more than 30,000 students pertaining to 7 distance learning courses, student demographics, course assessments and student interaction with virtual learning environment. We have merged tables together using unique identifiers. We will first explore the merged data using SAS to generate insights and then build appropriate predictive models.

Methods

Data Preparation and Analysis:

• Data were obtained by joining 7 different tables(Fig 1). The Student Info table contains demographic details of students, Student Registration contains information on when the students registered/unregistered for the courses, StudentVLE and VLE tables contain virtual learning environment information, Student Assessment, Assessment tables contain information on assessments.

• Information from VLE tables was summarized to get the total sum clicks for various types of activities the student undertakes for a course module. Each student undergoes several assessments over the duration of course. Assessments were weighted and students may opt to drop out of courses by withdrawing.

• The final dataset contains 26 variables with Final_result as the target.

• Fig 2. shows percentage rates of results by regions. South region had the highest pass percentage whereas Wales had the highest failure percentage and North Western region with highest percentage of withdraws.

• Fig 3. represents percentages of Pass, Fail and withdraws for each of the modules. Module AAA had the highest pass percentage and lowest percentage of failure whereas module CCC had the lowest pass percentage and highest percent withdraws.

• Fig4. provides frequencies of Pass, Fail and Withdraws for each assessment number.

• In Fig 5. Students who passed the modules had the most number of total clicks on materials compared to students who failed or withdrew from the courses.
**Decision Trees**

- Using decision tree, we would be able to explain the most important variables of our analysis by observing the top segment of the decision tree and analyzing the variable importance matrix.

- From fig 6. we can see that date_unregistration, latest_date_of_interaction, score, module and total_clicks were important variables and later we will use these to build the decision tree.

- Target variable like ours which was categorical required data to be partitioned, a chi square statistics was computed.

\[ x^2 = \sum \frac{e_i - n_i \hat{p}_i}{n_i} \]

\[-\log(p-value)\]

- The logworth statistic was used for pruning or growing a tree. The first split was done on the date_unregistration which split the data into two groups. Here this first split separates the 2 groups on 3 levels of the target variable.

- Other criteria used for splitting are Gini coefficient and Entropy.

- Gini coefficient identifies between the heterogeneous and homogenous groups.

- Fig 7. represents an interactive decision tree model which uses the classifying variables based on their (logworth) value.

- Fig 7 Decision Tree

- Misclassification rate of the train and validation data was close and was less than 0.10, so that was a fair estimate of a good model. However, The data split resulted into homogenous groups of withdrawn and pass students which at this point was not desirable.

- Hence we ignore this variables and instead do our first split on the second most important variable i.e latest_date_of_interaction.

- It split into two child nodes based on if the student had the latest interaction with the course material in greater than 206 days from the start of lesser. It gave us some interesting results which were aligned to previous research done of the similar subject. Students greater than 206 days had as low as 1.4% withdrawn percent rate.
Decision Trees (Continued)

<table>
<thead>
<tr>
<th>Obs</th>
<th>NAME</th>
<th>LABEL</th>
<th>RULES</th>
<th>IMPORTANCE</th>
<th>VIMPORTANCE</th>
<th>RATIO</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>latest_date_of_interaction</td>
<td>2</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.0000</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>score</td>
<td>3</td>
<td>0.3548</td>
<td>0.3515</td>
<td>0.9999</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>code_module</td>
<td>1</td>
<td>0.3895</td>
<td>0.1591</td>
<td>1.14409</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Agreement_rtm</td>
<td>3</td>
<td>0.3296</td>
<td>0.1394</td>
<td>1.07942</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Total_clicks</td>
<td>3</td>
<td>0.0796</td>
<td>0.0641</td>
<td>0.27592</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>date_submitted</td>
<td>1</td>
<td>0.0772</td>
<td>0.0772</td>
<td>1.01479</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>age_brad</td>
<td>2</td>
<td>0.0180</td>
<td>0.0128</td>
<td>0.51706</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>earliest_date_of_interacti</td>
<td>1</td>
<td>0.0890</td>
<td>0.0033</td>
<td>0.26166</td>
<td></td>
</tr>
</tbody>
</table>

Fig 10. Variable importance updated

- 2nd split was done on the score variable which is a cumulative of all the TMA (Tutor Marked Assessment) for that module per student. Group of students who scored greater than 57 had a high pass percentage of 90.01 as compared to the other group who scored less than 57.
- With increasing complexity of the tree, we needed to plot the tree size and variation explained at every level so we could find at which level the variation was minimum. Accordingly we pruned the tree to get the simplified version of the tree.
- Misclassification error for this model had similar values for Train and validation with a value of 0.18.

Conclusions

- The most important variables are latest_date_of_interaction that is measured as number of days relative to the start of module presentation, followed by Score, Code_module.
- Students who scored greater than 57 in initial assessments had a high pass percentage of 90.01 as compared to the other group who scored less than 57.
- Of the students who scored more than 73.5 and had total clicks greater than 1352 had the greatest pass percentage of 96.14 whereas only 3.4% ended up failing.
- Of the students who scored less than 73.5 and with clicks less than 1352, 8.16% ended up failing the course.

Future Work

- The scope of this project will be extended to do back test of the model and implement the successful validation results to identify students’ at risk, applicable to online course websites such as Coursera, Udacity.

References

- Kuzilek, J., Hlosta, M., Zdrahal, Z. Open University Learning Analytics Dataset. Data Literacy For Learning Analytics Workshop at Learning Analytics and Knowledge (LAK16)
- Decision Tree Models in Data Mining - Matthew J. Liberatore, Thomas Coghlan

Acknowledgement

- We thank Dr. Goutam Chakraborty, Professor, Department of Marketing, and founder of SAS and OSU Data Mining Certificate Program, Oklahoma State University for his constant guidance throughout the research.