OU Student Retention: Data exploration and Prediction

Springboard Data Intensive Capstone Final Report

By Soumya Krishnamurthy

Introduction

The initial idea for the project comes from thinking about motivation levels for online courses. In more detail, it leads to thinking about who are the students (gender and age distribution, background)?. How many of these students withdraw from the courses and is it possible to stop the trend?

The problem is to classify students who withdraw from distance learning courses so that earlier intervention can prevent them from withdrawing from the course

The client is Open University who can use the classification to decrease the proportion of students who withdraw from their courses. This can be applied to other courses too with a high degree of self-learning such as online courses. It will not only make the courses more relevant and useful to the students but also increase their success rate.

Background

Many studies have been conducted into student retention and progress at the Open University. Kennedy and Powell (1976) reported the following after conducting research on withdrawals from the Open University:

> Such students often have a demoralizing history of educational failure and bring feelings of insecurity and educational and intellectual inferiority to their studies. This problem is compounded by the fact that by joining the Open University a student is not simply put in contact with a body of knowledge; he is forced to accommodate himself to a specialized pattern of interaction and communication. The student does not only have to learn new vocabularies; he must learn to debate and communicate in a manner which is acceptable to the academic community (p. 69).

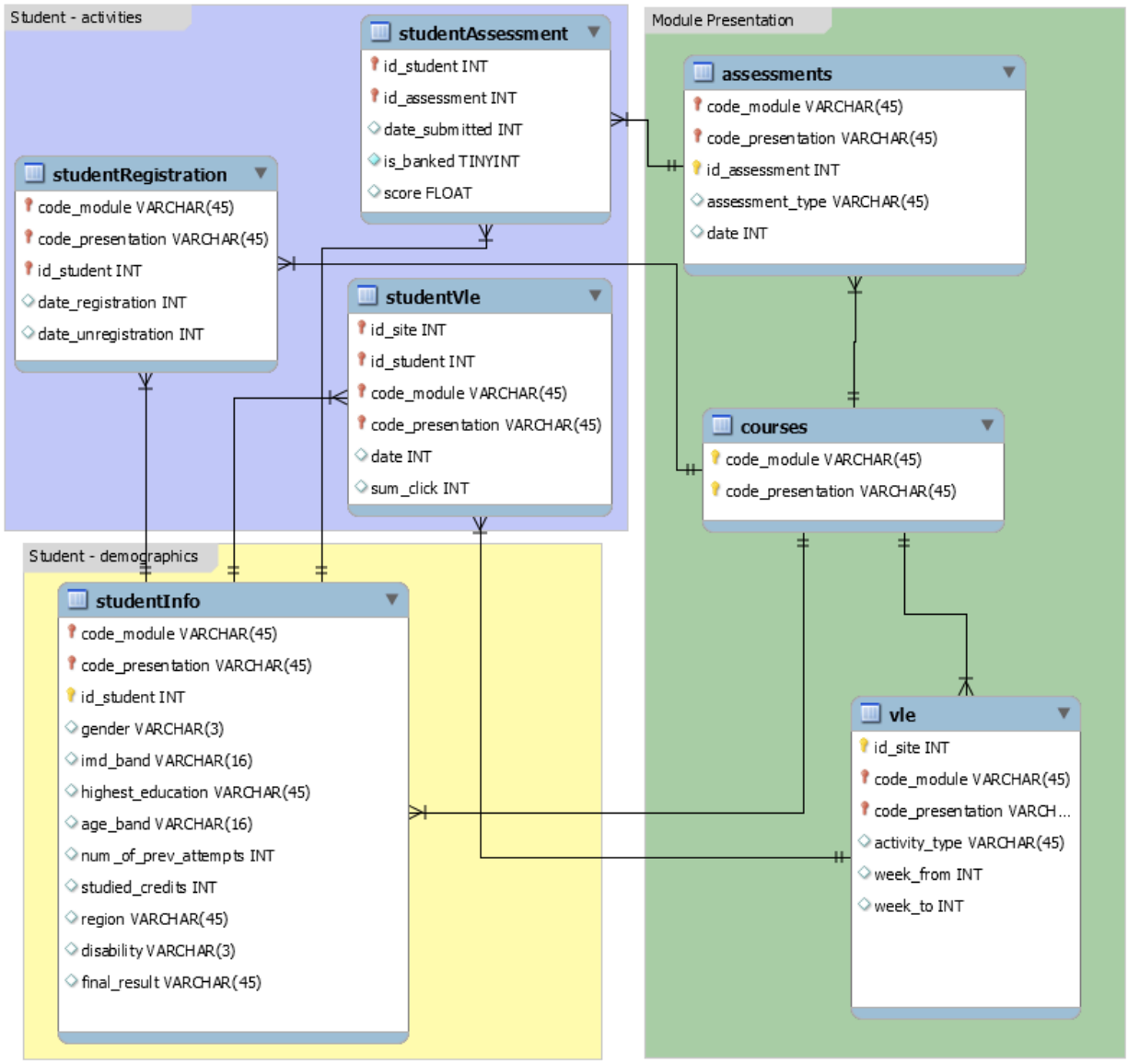
Simpson(2006) has reviewed ways to predict student success and highlighted the significance of statistical methods like logistic regression rather than assumptions and theories related to the profile of these students and influencing factors in their decision to withdraw. More recently, machine learning by Wolff and Zdrahal (2012) has been used to predict students who are at risk and struggling in class. This study uses the VLE engagement to tailor learning goals.

It is only possible to create the profile with limited success but nonetheless, the target profile can be used to intervene early on or even offer extra services to the students who are “at risk.” This analysis solution will provide a new insight based not on intuition but on concrete data analysis.

Data

The dataset contains multiple csv files giving a complete picture of the overall student data. Schema for the data can be found here:
https://analyse.kmi.open.ac.uk/open_dataset

The dataset contains information about the student profile(demographics), date of registration and unregistration, courses taken along with assessments, scores and vie engagement. Student withdrawal is a result of several personal, financial, psychological and other external factors. The information from the dataset is inadequate. In addition, it is a small subset of the overall data. Some missing information can also hinder analysis: Level of entry, course codes. Since the data was taken from a well established source, it was very clean with few missing values. The only complication I had was to integrate the data from multiple files to get the complete picture.



To start with, we look at the student profile and the possibility of predicting chances of withdrawal based solely on demographics. The data is in the file studentInfo.csv

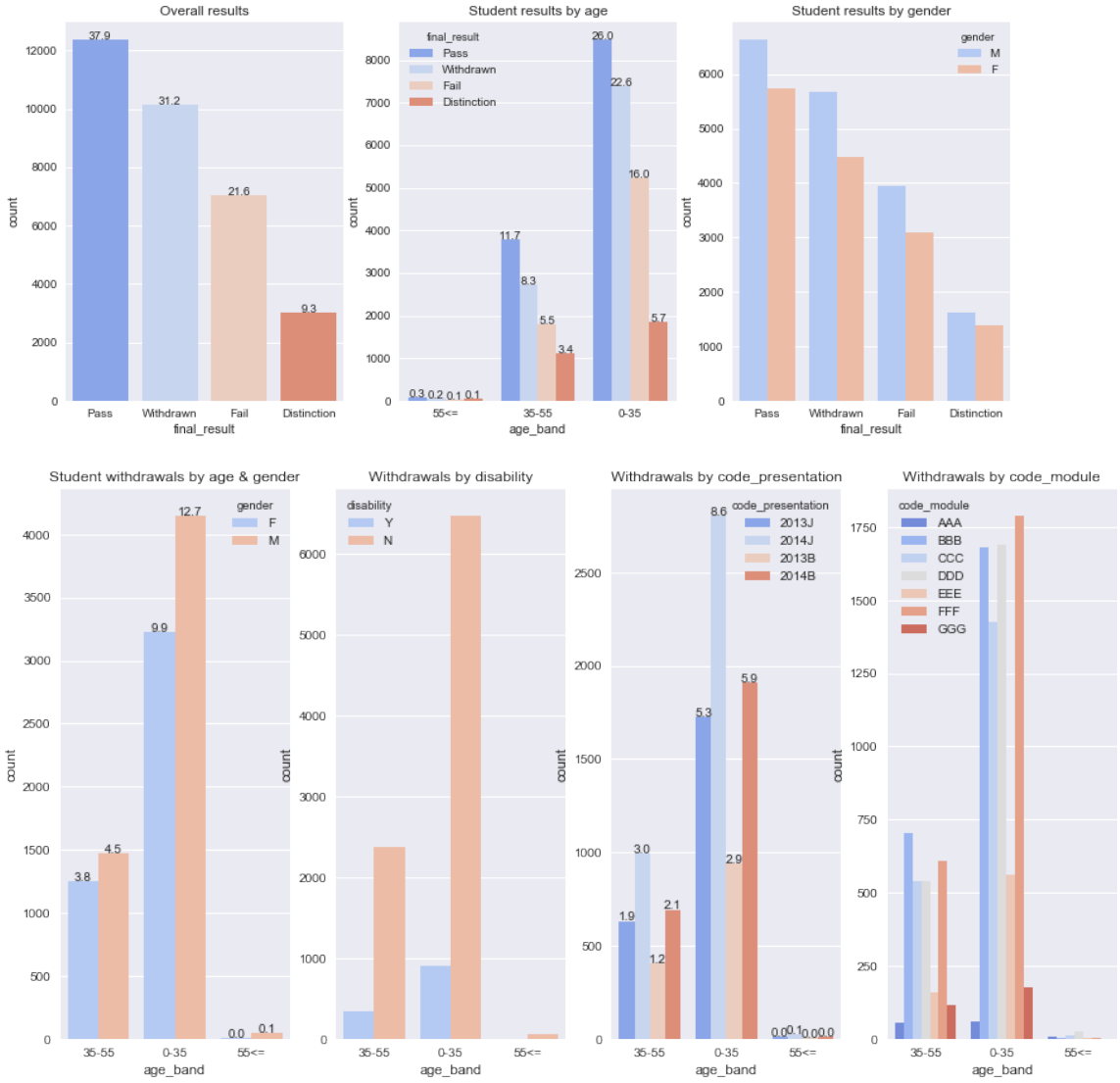
Column name	Description
id_student	Unique student identifier used across the database
gender, age_band and disability	Self-explanatory demographics
highest_education	highest student education level on entry to the module presentation.
imd_band	specifies the Index of Multiple Depravation band of the place where the student lived during the module-presentation
region	Identifies geographical region where the student lived while taking the module-presentation
num_of_previous_attempts	the number of times student has attempted this module
Studied_credits	total number of credits for the modules student is currently studying
code_modue	identification code for the module on which student is registered
code_presentation	code name of the presentation. It consists of the year and “B” for the presentation starting in February and “J” for the presentation starting in October.
final_result	student’s final result in the module-repsentation

Findings

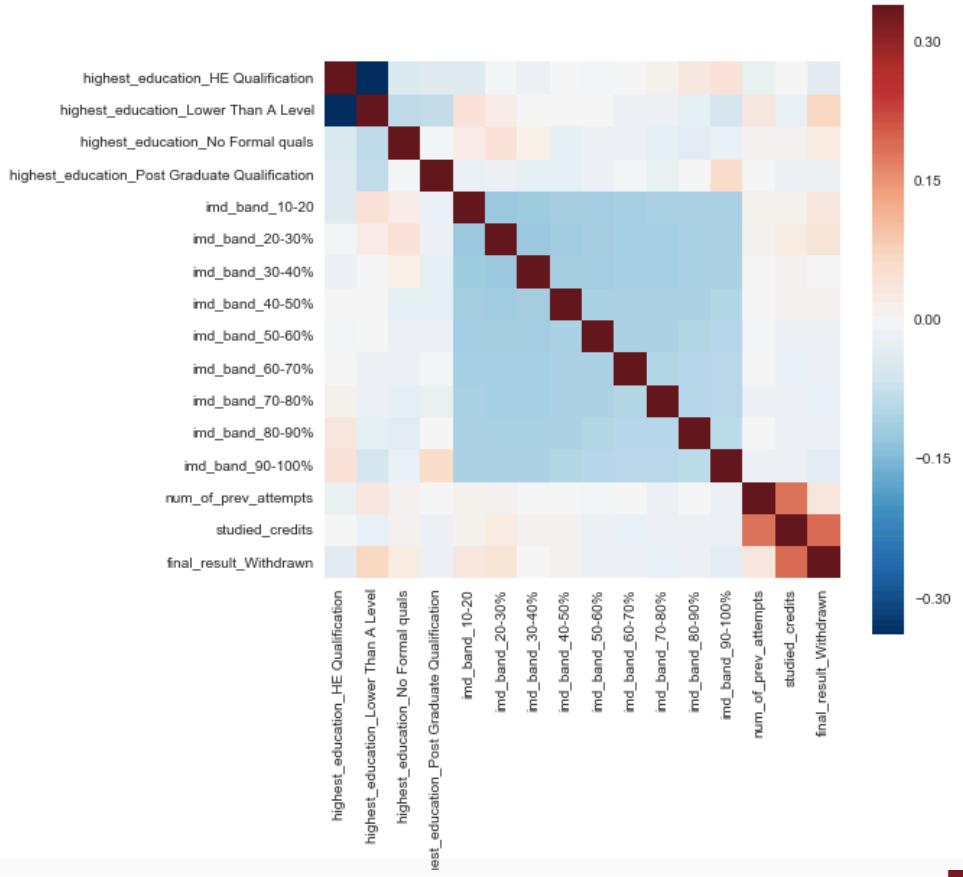
Part A

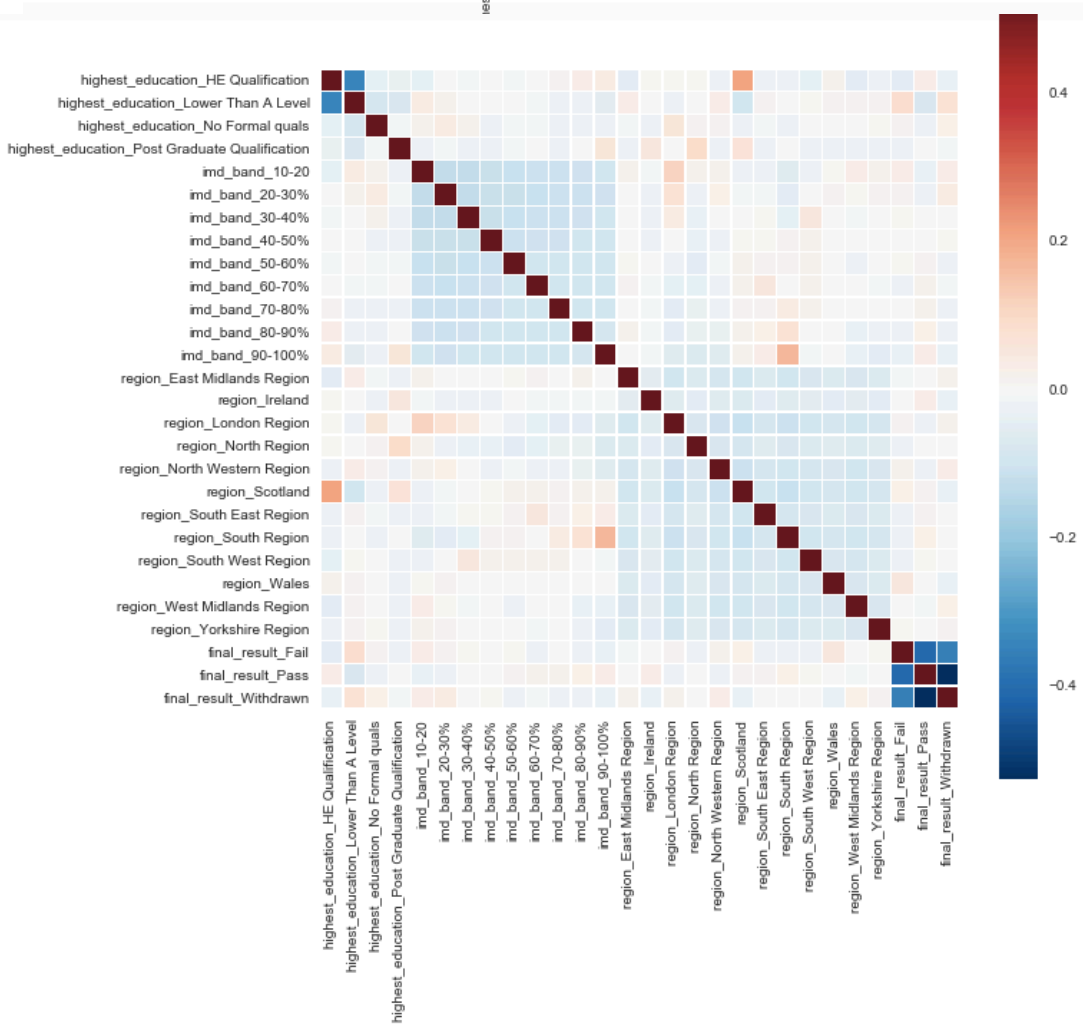
Exploratory data analysis based only on demographics. To gain an insight into the possible reasons the students withdraw, the aim is to find correlation between students who withdraw and:

* region
* imd_score
* highest education
* previous attempts
* number of credits



The initial findings about student result clearly show 31% (10156 out of 32593 records) withdraw from the course and this forms the target group. There is a fairly big increase in 2014 when compared to 2013. So there is a growing trend and there is a need to find solutions. To get a better idea of the correlations between the different factors, a heat map is used.

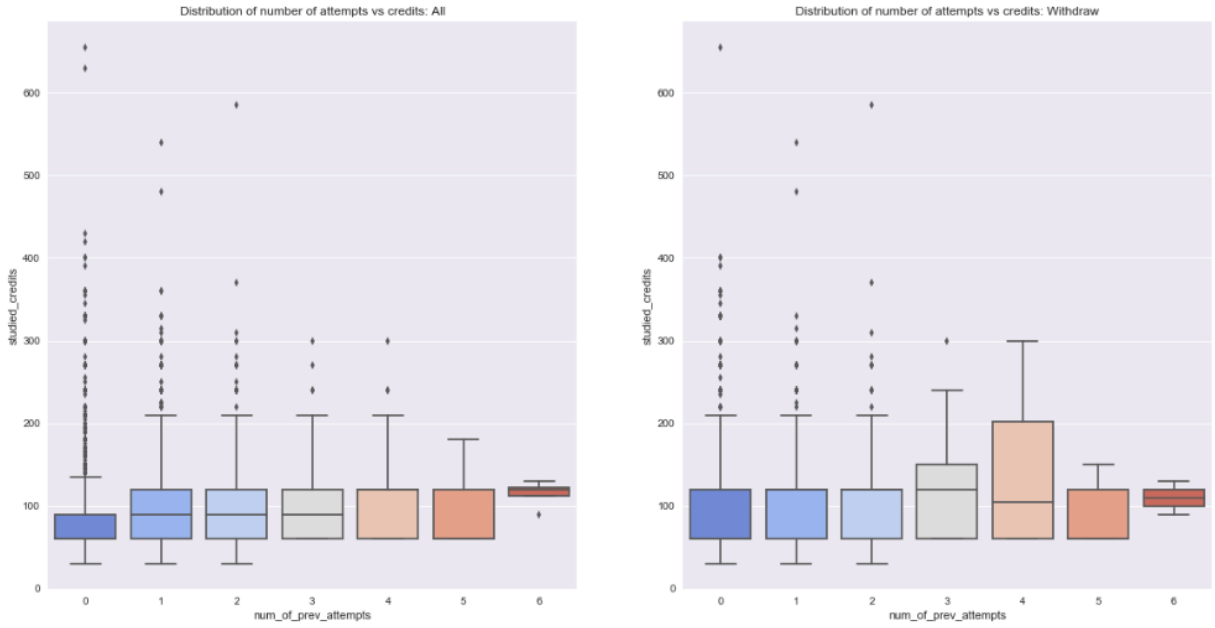

highest_education_HE Qualification
highest_education_Lower Than A Level
highest_education_No Formal quals
highest_education_Post Graduate Qualification
imd_band_10-20
imd_band_20-30%
imd_band_30-40%
imd_band_40-50%
imd_band_50-60%
imd_band_60-70%
imd_band_70-80%
imd_band_80-90%
imd_band_90-100%
num_of_prev_attempts
studied_credits
final_result_Withdrawn
0.30
0.15
0.00
−0.15
−0.30


highest_education_HE Qualification
highest_education_Lower Than A Level
highest_education_No Formal quals
highest_education_Post Graduate Qualification
imd_band_10-20
imd_band_20-30%
imd_band_30-40%
imd_band_40-50%
imd_band_50-60%
imd_band_60-70%
imd_band_70-80%
imd_band_80-90%
imd_band_90-100%
region_East Midlands Region
region_Ireland
region_London Region
region_North Region
region_North Western Region
region_Scotland
region_South East Region
region_South Region
region_South West Region
region_Wales
region_West Midlands Region
region_Yorkshire Region
final_result_Fail
final_result_Pass
final_result_Withdrawn
0.4
0.2
0.0
−0.2
−0.4

Exploratory data analysis

A lot of information can be gleaned from the initial data analysis. Some, even not directly related to the target, is fascinating.

1. Some regions have more students who withdraw than others. These are East and West midlands, London and the North.
2. Highest education achieved at the start of the course is very indicative: those with no Formal Quals or lower than A levels are more likely to withdraw than others.
3. imd_band is another important factor as those between 0-30% are most “at risk”.
4. The most positive correlation is studied_credits, which in turn, correlates with number of previous attempts. As we can see below, students who withdraw have more previous attempts and hence studied credits.



There may be other factors affecting the studied credits and pointers like assessments, courses taken and/or VLE engagement. That would form the basis for the next data story.**

PART B

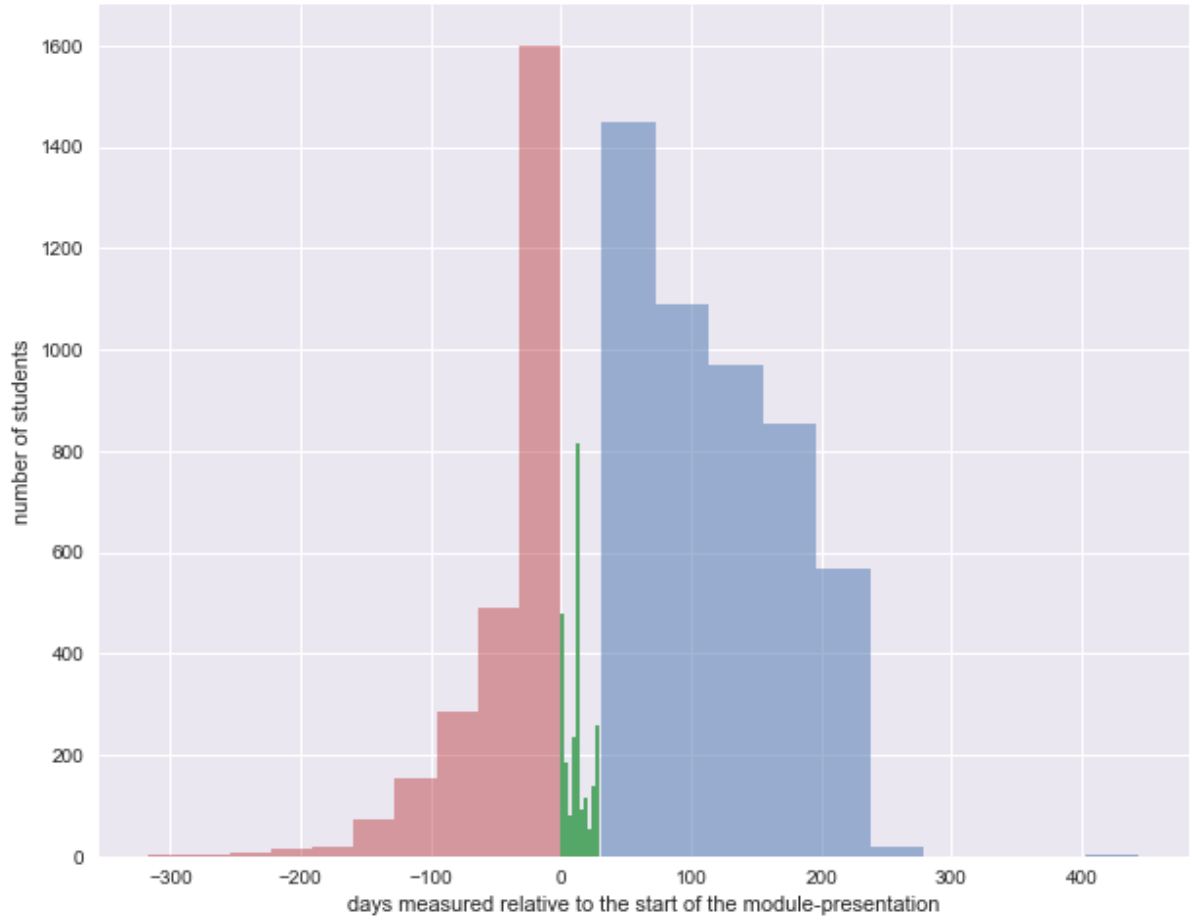
The next data story is done in 5 parts-correlation between student withdrawal and:

* date of withdrawal
* students who are disadvantaged due to their background
* scores and student withdrawal
* VLE engagement and student withdrawal
* disabled students analysis

Date of withdrawal:

The data is in the file: studentRegistration.csv. We are especially interested in the column date_unregistration – date of student unregistration from the module presentation (this is the number of days measured relative to the start of the module-presentation)
So a negative number implies students withdrew before the start of module and a positive number of say 35 that the student withdrew on the 35th day after the course started.

number of students who withdraw	10033	
number of students who withdrew before the start of module presentation	2643	26% of students withdrew before the start of the course, majority just before
number of students who withdrew in the first 30 days after the start of the module presentation	2446	25% withdraw in the first month
mean module presentation length (days)	255.55	



Disadvantage due to background

As we have observed before:

- More students who withdraw have A Level or Lower Quals
- They are more likely to be in 0-30% imd_band

We define these as disadvantaged students or those most at risk.

Number of students from A level or lower Quals: 13505		
Withdrawn	4769	47.5% of all withdrawals
Pass	4472	
Fail	3521	
Distinction	743	
Number of students from lower imd_band(0-30%): 6965		
Withdrawn	2552	25% of all withdrawals
Pass	2222	
Fail	1760	
Distinction	431	

Number of students deemed at risk as they have educational or imd_band disadvantage:		52.5% of 32593 students
Withdrawn	6006	60% of all withdrawals
Pass	5774	
Fail	4299	
Distinction	1048	

```
previous attempts: disadvantaged students
0    14721
1     1867
2      406
3       94
4       26
5       10
6        3
Name: num_of_prev_attempts, dtype: int64
------------------------
previous attempts: all students

0    28421
1     3299
2      675
3      142
4       39
5       13
6        4
Name: num_of_prev_attempts, dtype: int64
```

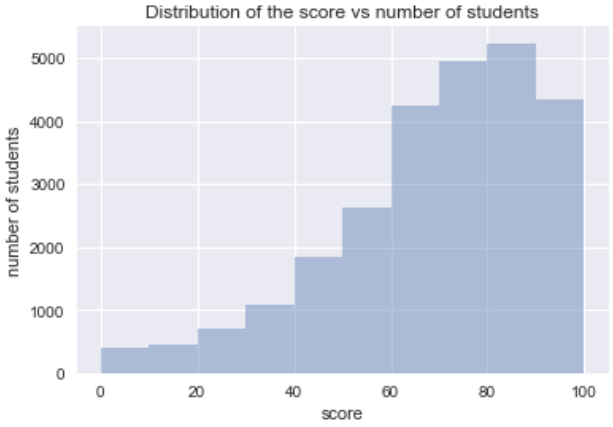
Out of 32593 enrolled students, 17127 or 52.5% deemed at risk as they have both educational as well as imd_band disadvantage
Among these 17127 students, 6006 withdrew or 35%. However, out of 10033 students who withdrew overall, 60% are in the disadvantaged category. We had previously noticed that number of previous attempts had the most positive correlation. This is observed to the right.

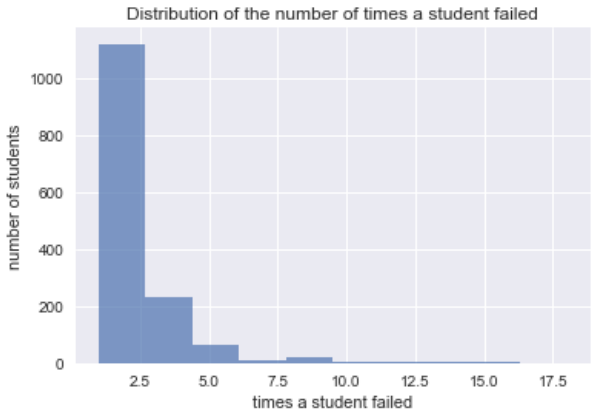
To predict the probability that a student will withdraw, we will use the above information.

Scores and withdrawal

Many attempts in the same module can either mean :
- Student failed multiple times
- Student took the course but did not complete assessments

The file assessments.csv provides data related to assessments taken by student as well as score. Analyzing student scores can be a project by itself but in the data provided, it is missing. It is mentioned as future work. So the focus is on the times a student has failed and if this is a contributing factor to withdraw.





The scores for students who withdrew appears to be good. Many students appear to have failed fewer than twice leading us to conclude that the score is not as related to withdrawal as simply taking the assessments. Of the total, only 12.5% of the scores belong to students who withdrew.

VLE engagement and student withdrawal

VLE engagement data can be found in the file studentVle.csv. The only data provided is the sum_clicks, total number of clicks by a student. So, the analysis just involved finding the totals for all students vs those who withdrew.

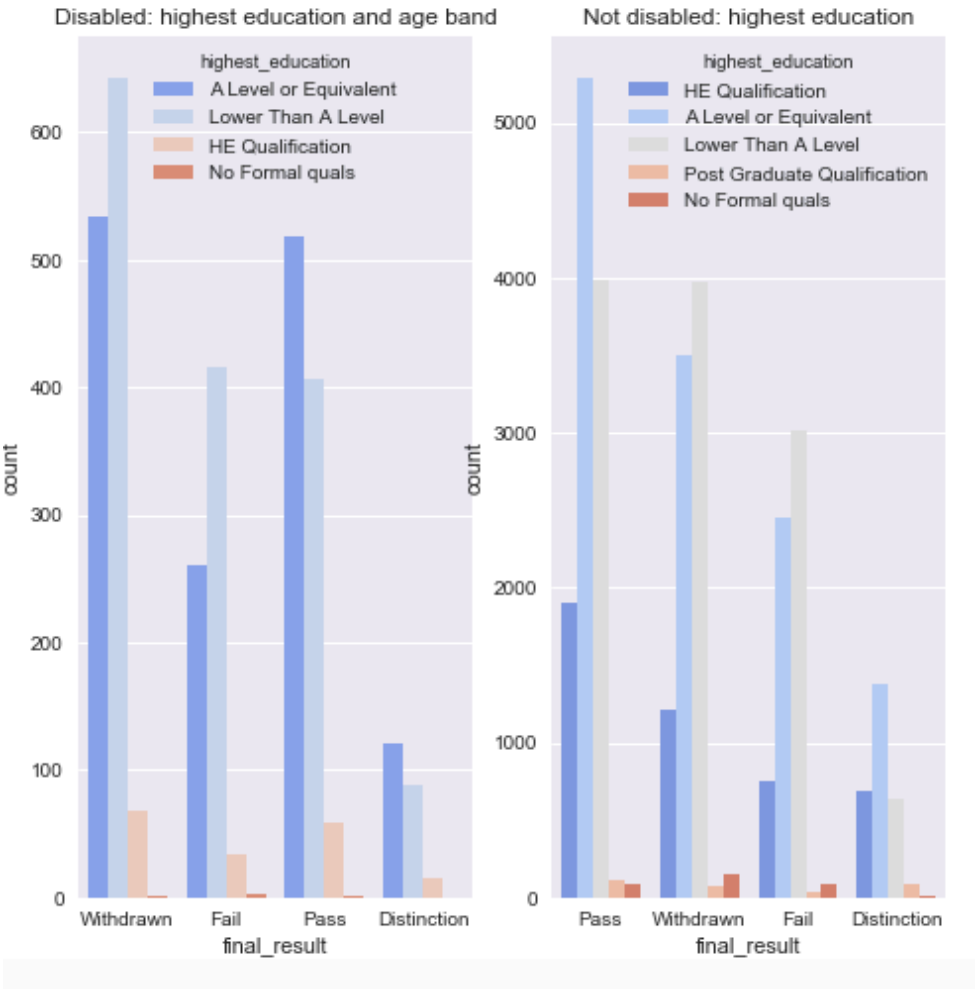
As we can see from the descriptive stats, the difference between all students vs the ones who withdrew is not vast. We **can** be sure that more VLE engagement will have a positive impact on students experience. Further work can be done to see if this has any direct impact on scores.

The data for analysis in depth is missing in this dataset. But lack of VLE engagement does not appear to be a direct reason for withdrawal.

```
(1830536, 17)
Unique id: 6769
All Students
count    1.300658e+07
mean     3.706290e+00
std      8.962795e+00
min      1.000000e+00
25%      1.000000e+00
50%      2.000000e+00
75%      3.000000e+00
max      6.977000e+03
Name: sum_click, dtype: float64
-------------------------
Students who withdraw
count    1.830536e+06
mean     3.485032e+00
std      8.405775e+00
min      1.000000e+00
25%      1.000000e+00
50%      2.000000e+00
75%      3.000000e+00
max      3.958000e+03
Name: sum_click, dtype: float64
```

Disabled students analysis

Lastly, more out of curiosity than evidence from previous data analysis, we look into disabled student data. The results are encouraging. Open University is the preferred choice for students with both physical or mental disabilities (OU, 2017). The final result based on for disabled students is summarized below:



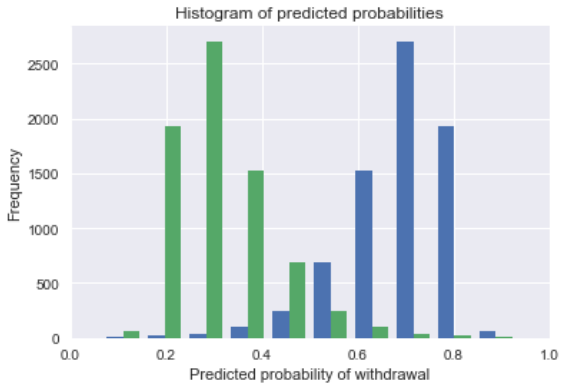
The disadvantaged students with A level or Lower withdraw more than any other group. With more data, it is possible to look at student retention as a separate topic.

Prediction

As mentioned in the Introduction, the target of his study is to predict a binary response based on a set of explanatory discrete values. The best suited for this kind of classification is Binary Logistic Regression. Our dependent variable or output is the probability of Withdrawal and the independent variables are the factors highlighted as “high-risk”. The probability can be used to intervene to retain the student. 10-fold cross validation is used to verify the result.

We can predict if a student will withdraw with up to 69% accuracy.

```
Data Prep
Training
Predicting
[False False False ..., False False False]
class probabilities: [[ 0.70219784  0.29780216]
 [ 0.6784059   0.3215941 ]
 [ 0.66707329  0.33292671]
 ...,
 [ 0.69715898  0.30284102]
 [ 0.75905892  0.24094108]
 [ 0.58609081  0.41390919]]
```




```
generating metrics
accurancy score: 0.695234199223
roc score: 0.631720450985
f1 score: 0.614897724049
classification report:
             precision    recall  f1-score   support

          0       0.70      0.96      0.81      6741
          1       0.55      0.10      0.17      3037

avg / total       0.66      0.70      0.61      9778

confusion matrix
[[6482  259]
 [2721  316]]
```

To understand the result better, we can use evaluation metrics. In particular, the target was to get more than 60% as:

- accuracy score
- roc_score
- f1_score
- cross validation score (10-fold)
- confusion matrix

As we can see from the metrics, there are two glaring errors:
Confusion matrix shows non-withdrawals have been predicted much better than withdrawals (f1 score illustrates this even better) which is wrong
The precision and recall score do not match showing it is a poor classifier.

The main reason for this is the imbalance of data as most of the data relates to the students who did not withdraw (70%). To correct this, we use over-sampling using Synthetic Minority Over sampling Technique (SMOTE)

Results from over-sampling and using balanced class weights in Logistic Regression

```
generating metrics
accurancy score: 0.602317462675
roc score: 0.636845831966
f1 score: 0.579154221034
classification report:
                   pre       rec       spe        f1       geo       iba       sup

          0       0.59      0.66      0.55      0.62      0.60      0.36      6733
          1       0.61      0.55      0.66      0.58      0.60      0.36      6730

avg / total       0.60      0.60      0.60      0.60      0.60      0.36     13463

confusion matrix
[[4425 2308]
 [3046 3684]]
Mathews Correlation 0.205856408574
Cohens kappa 0.204615454592
The geometric mean is 0.5997974415488487
```

This is much better. We can see that the precision and recall scores match and the confusion matrix correctly predicts both withdrawals and non-withdrawals. This is seen in the f1 scores too. But the accuracy is only 60%.
A final attempt will be made to use the over-sampled data with other classifiers, particularly Random Forest and kNN to see if better accuracy can be obtained.

Results from using Random Forest

```
For Random Forest Classifier:
[ 0.68777849  0.66613622  0.67695735  0.67027371  0.67971983  0.68194842
  0.6768545   0.68184713  0.68407643  0.67579618] 0.67813882757
generating metrics
The geometric mean is 0.6731067681804715
[[4725 2008]
 [2385 4345]]
                   pre       rec       spe        f1       geo       iba       sup

          0       0.66      0.70      0.65      0.68      0.67      0.45      6733
          1       0.68      0.65      0.70      0.66      0.67      0.46      6730

avg / total       0.67      0.67      0.67      0.67      0.67      0.45     13463

accurancy score: 0.673698284186
f1 score: 0.664220744478
Mathews Correlation 0.347934499234
Cohens kappa 0.347388391481
```

This is even better. We have a much Bettie Cohen's kappa score improved too. For now this is the best result. Lets compare with kNN just to check if a better result can be obtained.

Result using kNN

```
For K-Nearest Neighbors Classifier:
[ 0.62635264  0.62539784  0.62412476  0.62189688  0.63737663  0.63642152
  0.62718879  0.63535032  0.62993631  0.63184713] 0.629589282446
The geometric mean is 0.6300833226485281
[[4171 2562]
 [2417 4313]]
                   pre       rec       spe        f1       geo       iba       sup

          0       0.63      0.62      0.64      0.63      0.63      0.40      6733
          1       0.63      0.64      0.62      0.63      0.63      0.40      6730

avg / total       0.63      0.63      0.63      0.63      0.63      0.40     13463

Mathews Correlation 0.260407096397
```

Better than Logistic Regression but nor Random Forest

Recommendations

The following recommendations are meant to intervene once a student registers for a course to retain he student and increase success rate by reducing withdrawal rate. The interventions may be (A) reactive, or (2) proactive. I

1: Proactive: Many students at risk withdraw even before the course starts or in the first month. This might mean the students are not sure they are ready for the course. So r**egistration should be combined with counseling.** This can include matching students with appropriate courses (data connecting withdrawal related to courses will be very useful), raising awareness about expected work loads and adding a personal touch to distance learning.

2: Since a majority of students who withdraw do not have A levels, a summer or **pre-course preparation** for "high risk" students could calm the nerves and also enable them to start on a positive note.

3: **Tutorials and peer-peer engagement** to motivate the students and make sure the next two recommendations are followed.

4: Reactive: **Monitor VLE engagement**, especially in the first month and when it is low, try to understand reasons and offer solutions. This could also mean the course chosen is not interesting to the student.

5: Reactive: **Monitor assessment scores** to evaluate student response to materials, their strengths and weakness, adapt assessment.

6: With more data related to about course and assessment, a **personalized course** can be offered to students “at risk”.

Limitations and Further Research

The most important limitation was incompleteness of data. In particular, data related to course details, assessment and VLE engagement. Secondly, as mentioned earlier, student demographic can only predict chances of withdrawal to some extent. The dataset is a very small subset and hence, the validity of the findings depends on the reliability of data.

Further research can include:

- Personalized courses
- Profile by region
- Tracking student progress
- VLE engagement and its relation to scores

From a machine learning perspective, it has been observed that Random Forest gave better result than kNN or Logistic Regression. The only sampling technique used is SMOTE. In both cases, more can be done. Use other sampling memos as SMOTE has the disadvantage of overfitting. A mix of under- and over- camping is one way to go. More classifiers can be used to see if a better accuracy can be obtained.

Conclusion

Student retention in distance learning courses like the Open University, requires a different approach when compared to traditional university structure. This capstone used freely available data from OU Analytics to get an insight into the demographic profile of students who withdraw from the course. While it is not possible to analyze in depth, it is possible to identify the group most “at risk" as well as the date when students are more likely to withdraw. The profile is based on highest education received before the start of course (Lower than A level has a higher incidence) and imd-band (lower the score, greater the risk). This information allowed us to apply classifiers like Logistic Regression, Random Forest and kNN to predict with 68% accuracy the probability of a student withdrawing from the course and we could offer some recommendations. Further research can enable intervention to be more targeted to offer personalized solutions and increase student retention.

References

Open University (2017) http://www.open.ac.uk

Kennedy D. and Powell R. (1976) Student progress and withdrawal in the Open University. *Teaching at a Distance* 7 (November), 61–75.

Simpson, O. (2006). Predicting student success in open and distance learning. Open Learning, 21(2), 125-138.

Wolff , A. and Zdrahal, Z. (2012). Improving retention by identifying and supporting "at-risk" students. EDUCAUSE Review Online