OU Capstone

Soumya Krishnamurthy

6 June 2017

Introduction

- Client: Open University, UK (distance learning)
- Problem: 30% students withdraw from the course
- Approach:
    - exploratory analysis: most important factors
    - in depth analysis into these
    - in addition, brief insight into scores, VLE engagement and if they influence the decision
- Predict whether student will withdraw or not

Background

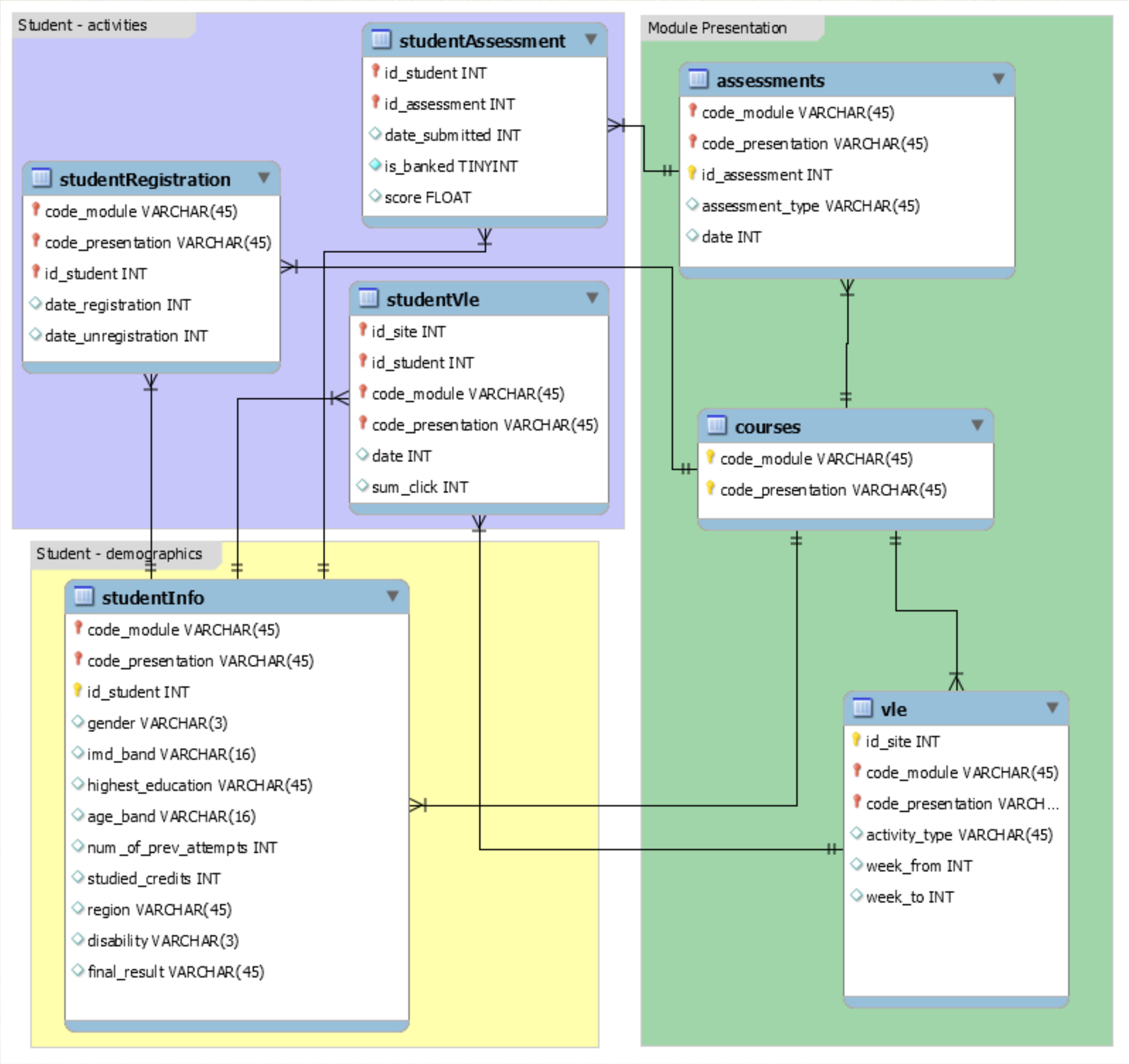
- Kennedy and Powell (1976): insecure, educational history of failure, not prepared to take the course

- Simpson (2006): Reviews ways to predict student success and significance of statistical methods like Logistic regression

- Wolff and Zdrahal (2012): uses machine learning to forecast if student is at risk of failing based on VLE engagement

Data

- Data is from website:
  https://analyse.kmi.open.ac.uk/open_dataset
  7 files in all providing information about:
  - Student demographics
  - Module/course information including assessment scores and VLE engagement
  - student activities like unregistration, scores and number of clicks on VLE
- Most important target file: studentInfo.csv related to demographics



Student - activities

studentAssessment
- id_student INT
- id_assessment INT
- date_submitted INT
- is_banked TINYINT
- score FLOAT

studentRegistration
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_student INT
- date_registration INT
- date_unregistration INT

studentVle
- id_site INT
- id_student INT
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- date INT
- sum_click INT

Student - demographics

studentInfo
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_student INT
- gender VARCHAR(3)
- imd_band VARCHAR(16)
- highest_education VARCHAR(45)
- age_band VARCHAR(16)
- num_of_prev_attempts INT
- studied_credits INT
- region VARCHAR(45)
- disability VARCHAR(3)
- final_result VARCHAR(45)

Module Presentation

assessments
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)
- id_assessment INT
- assessment_type VARCHAR(45)
- date INT

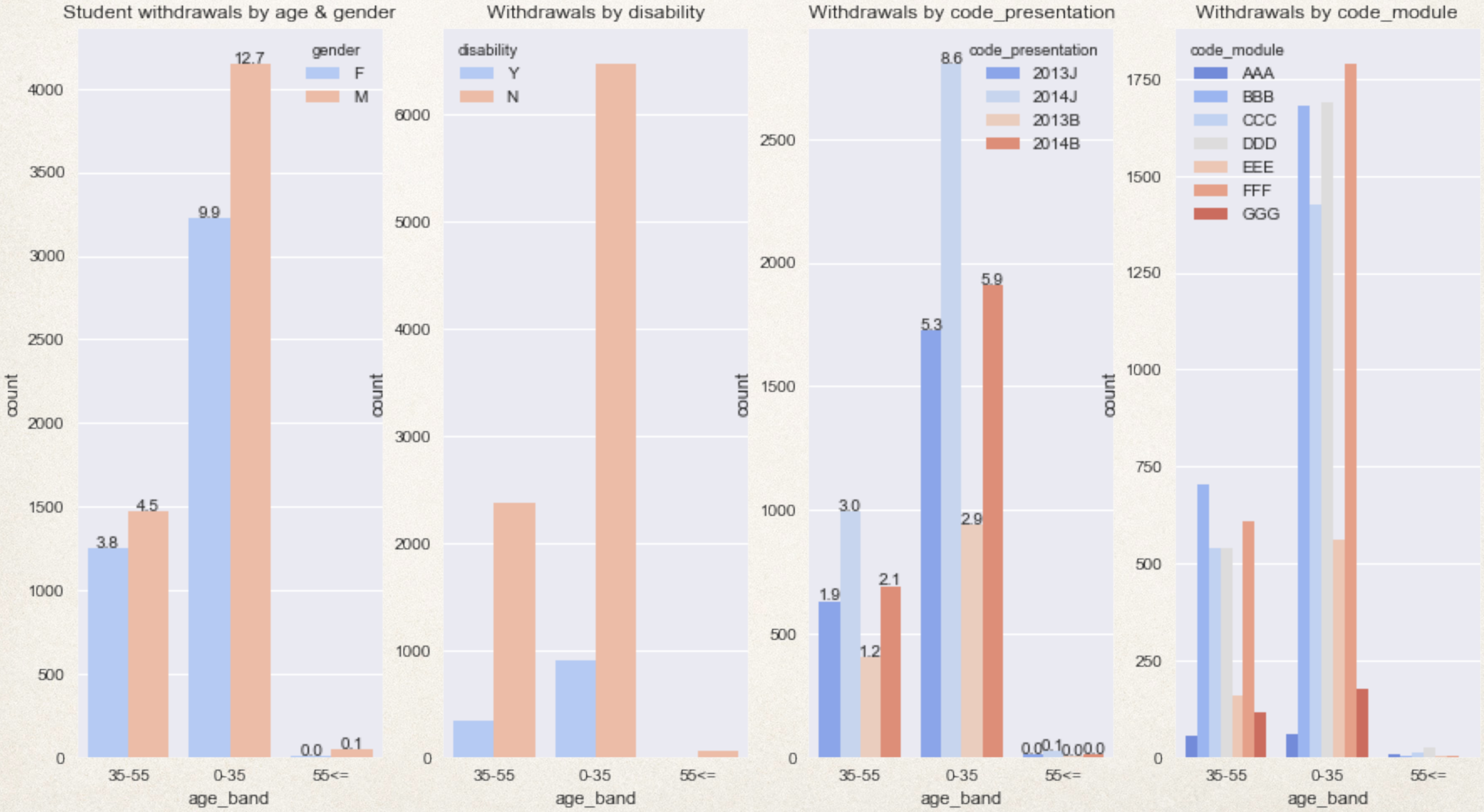
courses
- code_module VARCHAR(45)
- code_presentation VARCHAR(45)

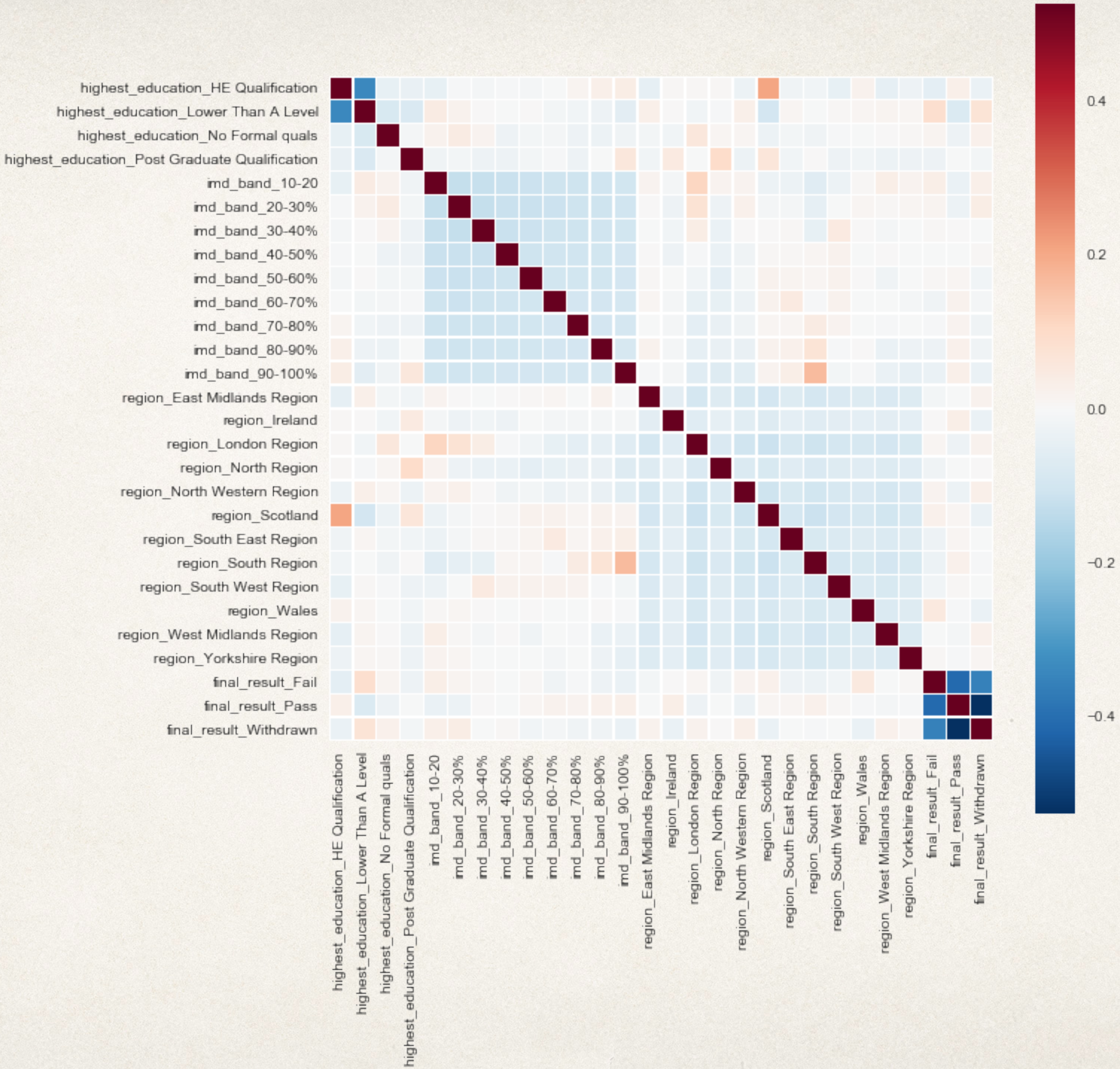
vle
- id_site INT
- code_module VARCHAR(45)
- code_presentation VARCH...
- activity_type VARCHAR(45)
- week_from INT
- week_to INT

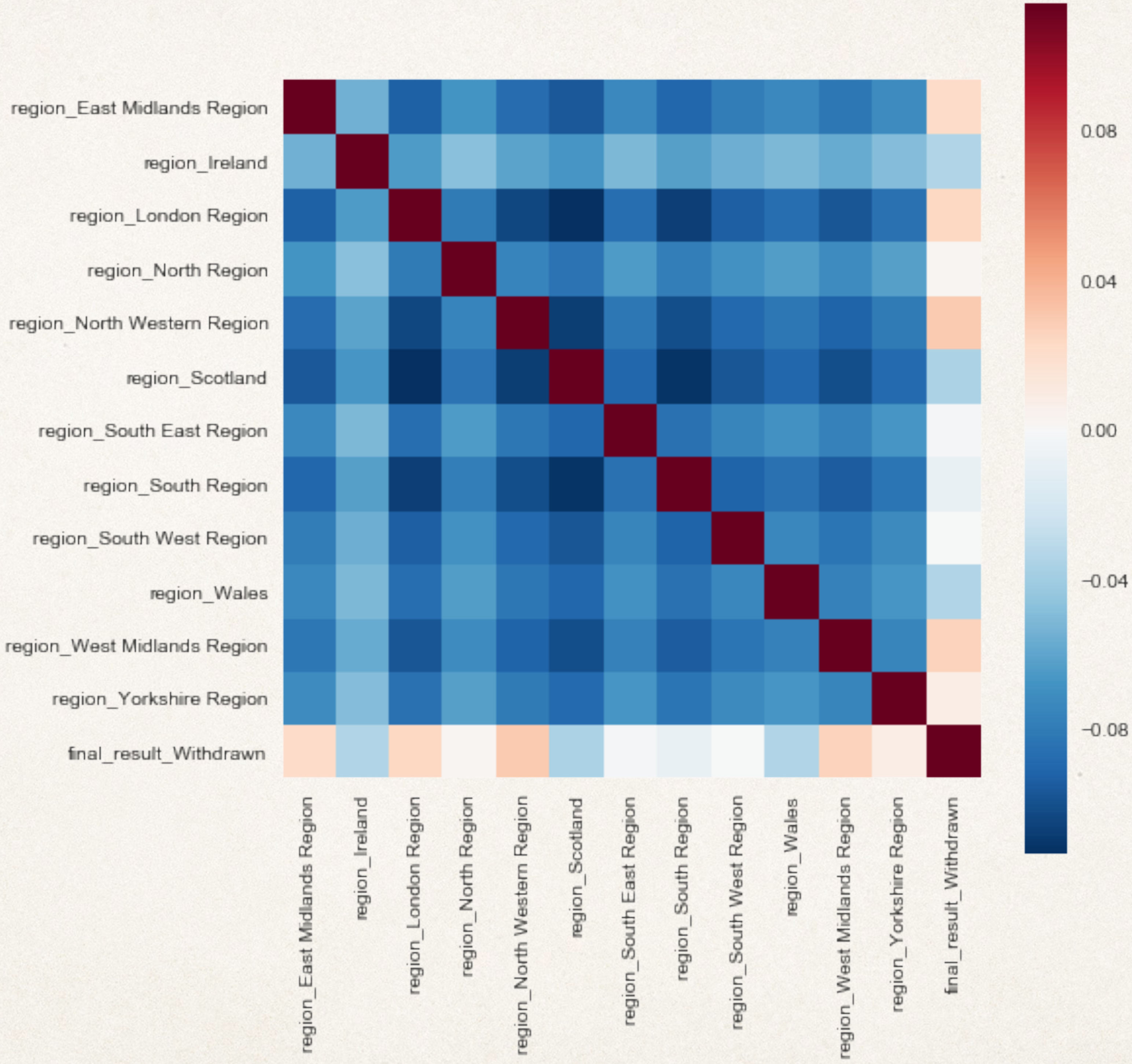
Exploratory data analysis based only on demographics

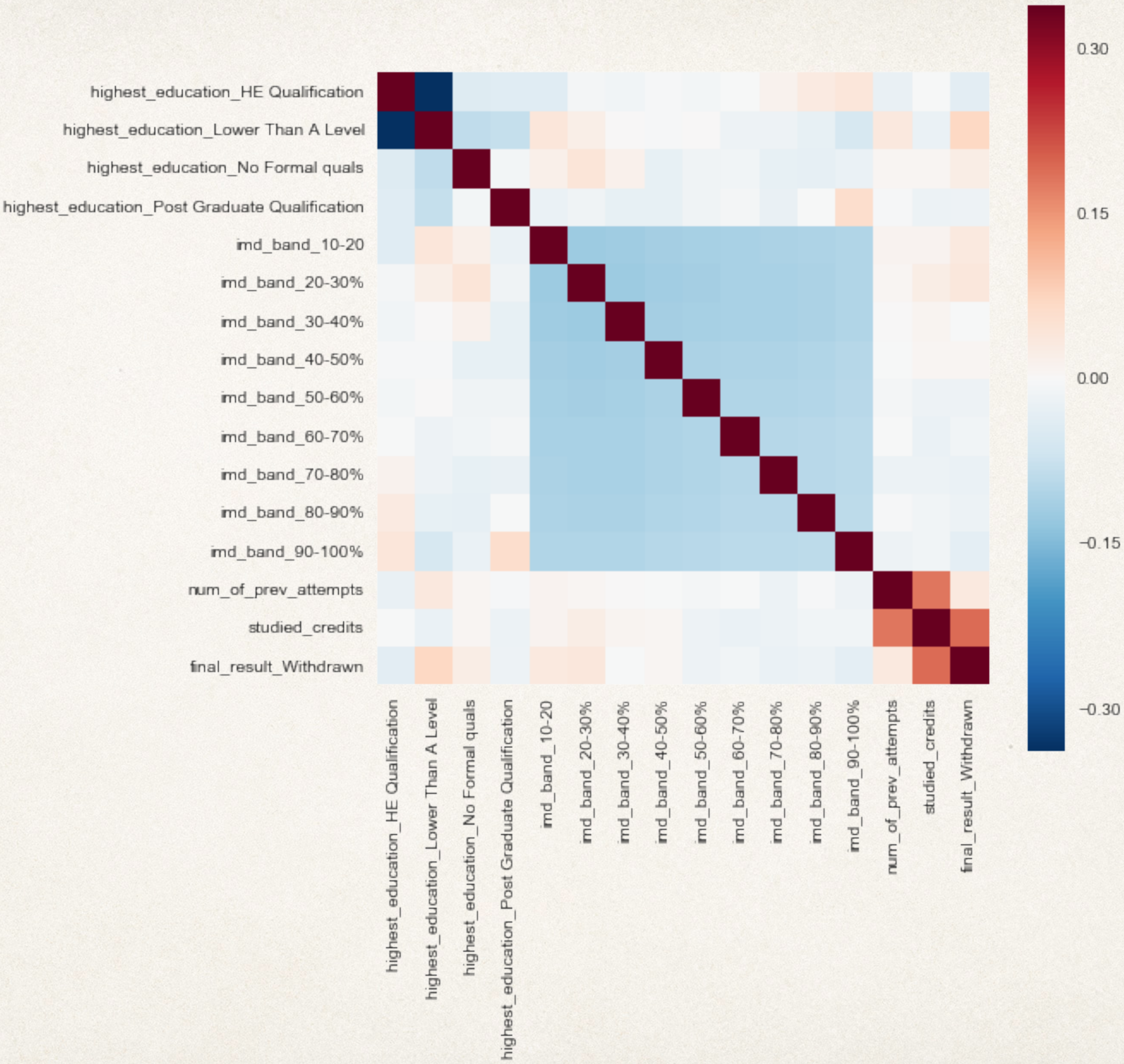
To gain an insight into the possible reasons the students withdraw, the aim is to find correlation between students who withdraw and:
* region
* imd_score
* highest education
* previous attempts
* number of credits

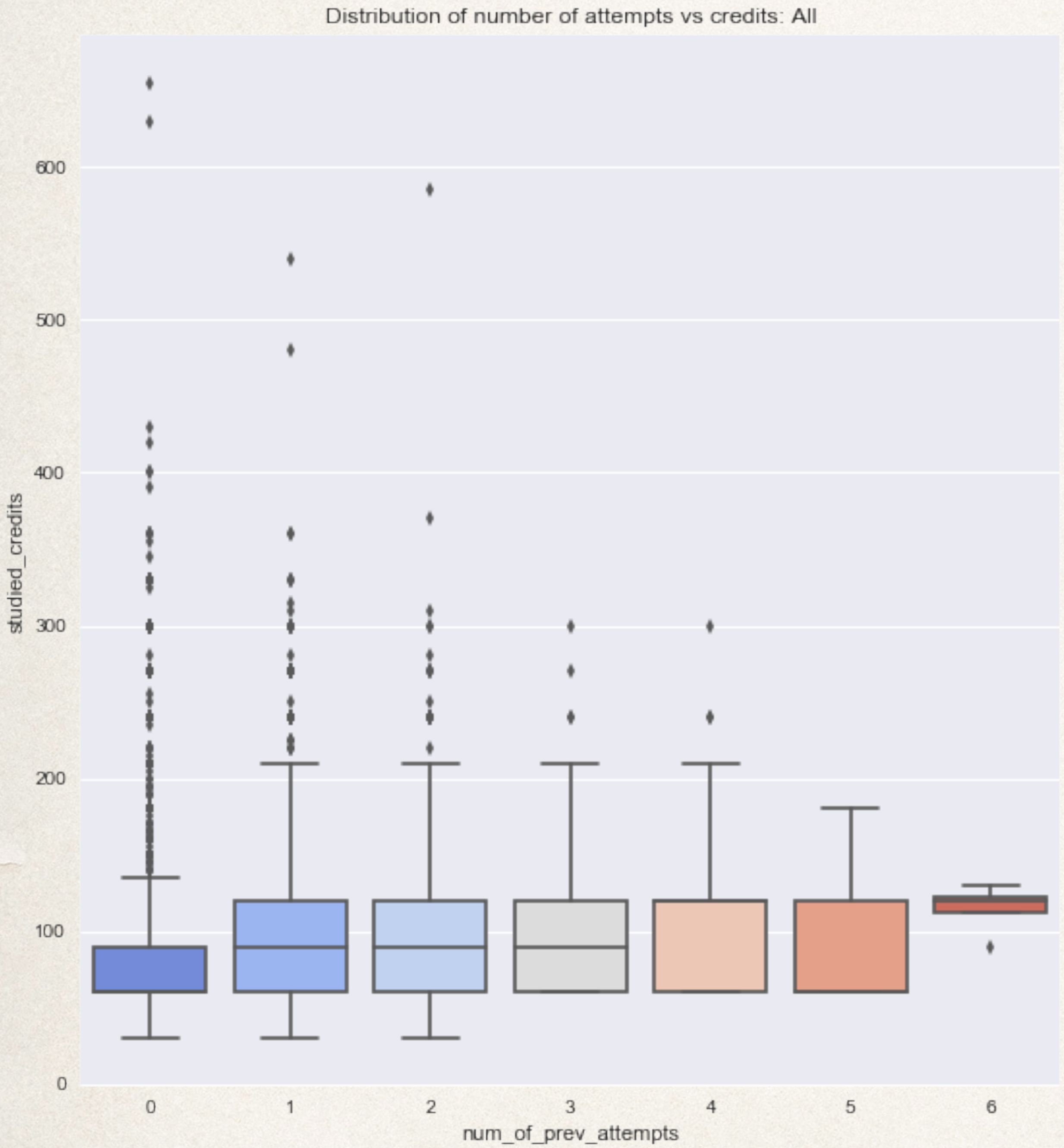
Overview of student data

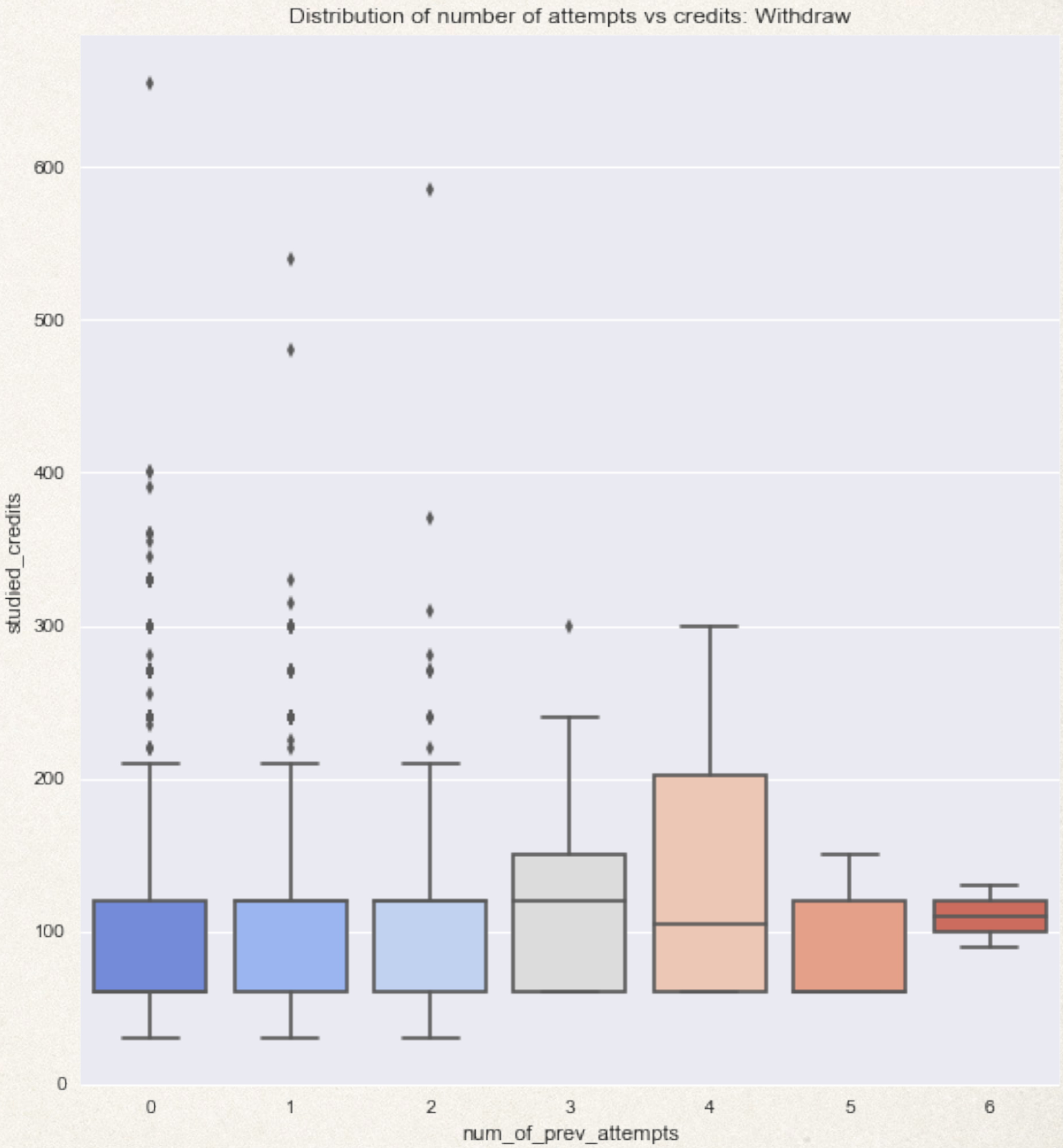


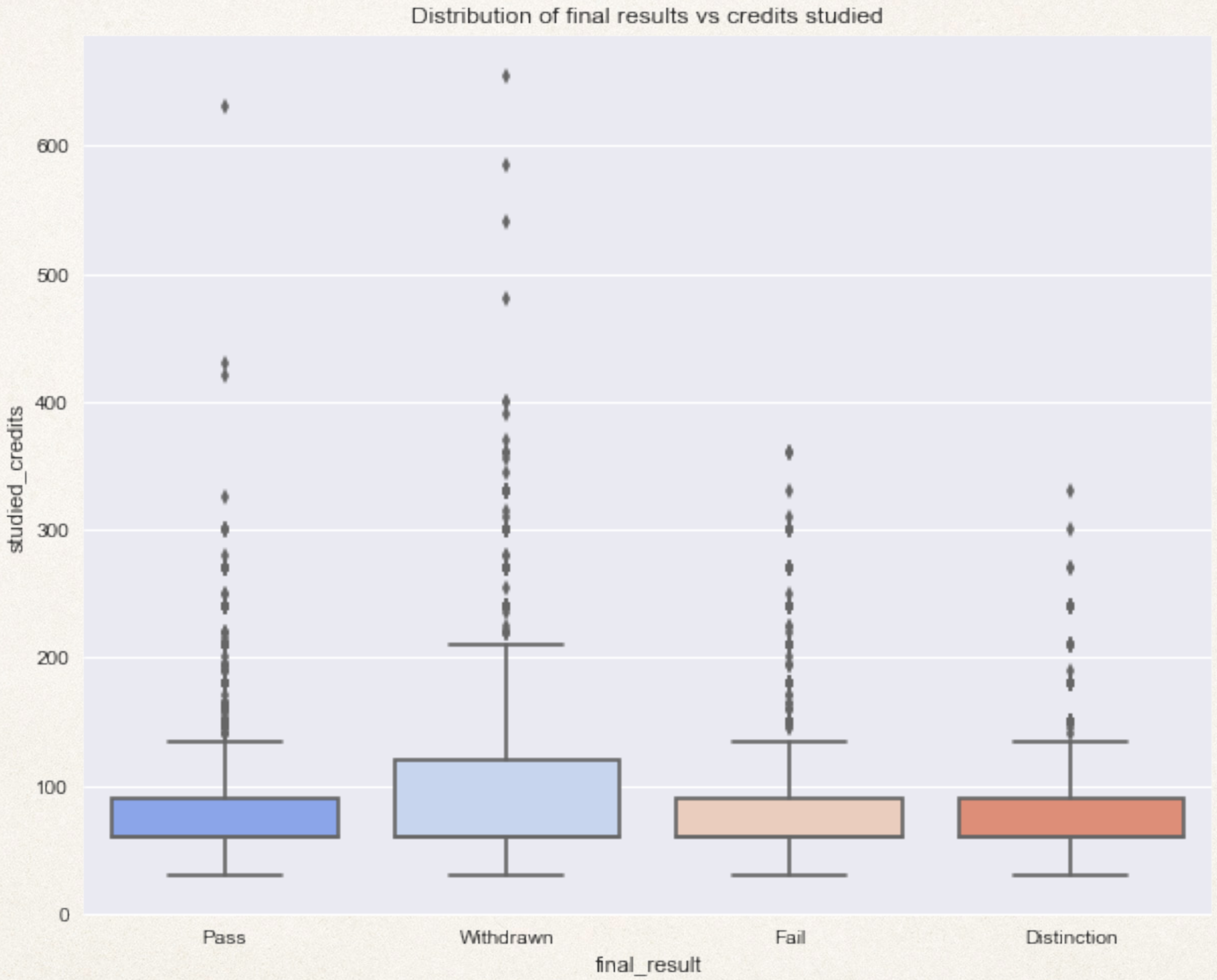

highest_education_HE Qualification
highest_education_Lower Than A Level
highest_education_No Formal quals
highest_education_Post Graduate Qualification
imd_band_10-20
imd_band_20-30%
imd_band_30-40%
imd_band_40-50%
imd_band_50-60%
imd_band_60-70%
imd_band_70-80%
imd_band_80-90%
imd_band_90-100%
region_East 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
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


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_Withdrawn
0.08
0.04
0.00
−0.04
−0.08


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


Distribution of number of attempts vs credits: All
studied_credits
0
100
200
300
400
500
600
num_of_prev_attempts
0
1
2
3
4
5
6


Distribution of number of attempts vs credits: Withdraw
studied_credits
0
100
200
300
400
500
600
num_of_prev_attempts
0
1
2
3
4
5
6


Distribution of final results vs credits studied
studied_credits
0
100
200
300
400
500
600
Pass
Withdrawn
Fail
Distinction
final_result

Initial findings

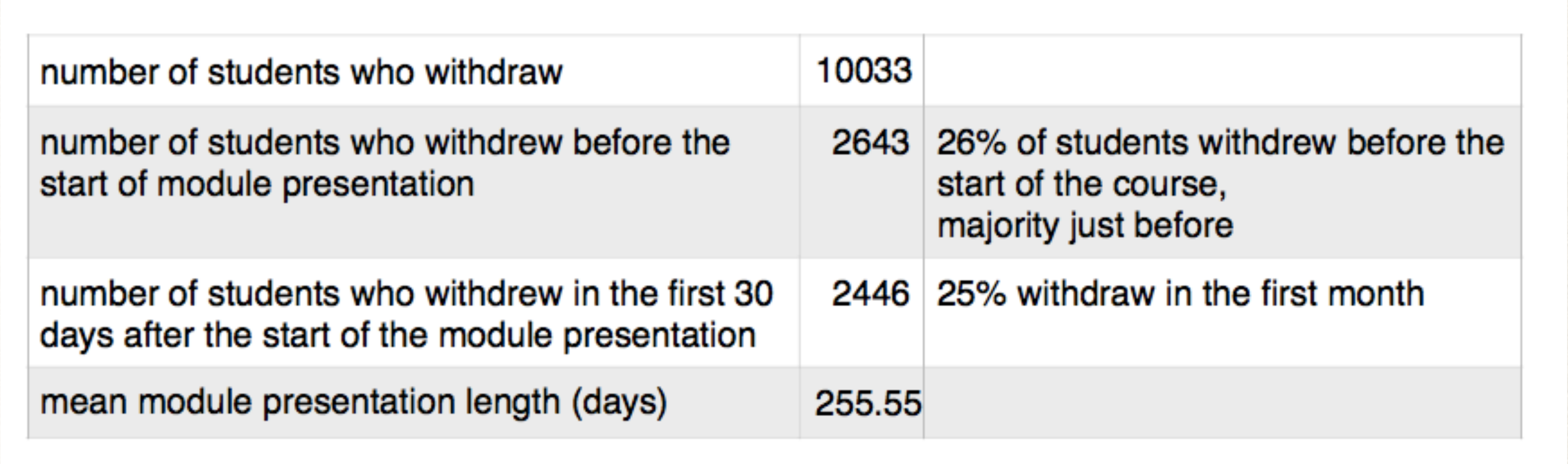
- region: positive correlation from East and West Midlands, London and North-West
- imd_score: more withdrawals from those in 0-30%
- highest education: Lower than A level or No Formal Quals has an impact
- previous attempts: More the number of attempts, especially more than 3, has a positive correlation
- studied credits: only students who withdraw have upper quartile credits more than 100

In-depth analysis

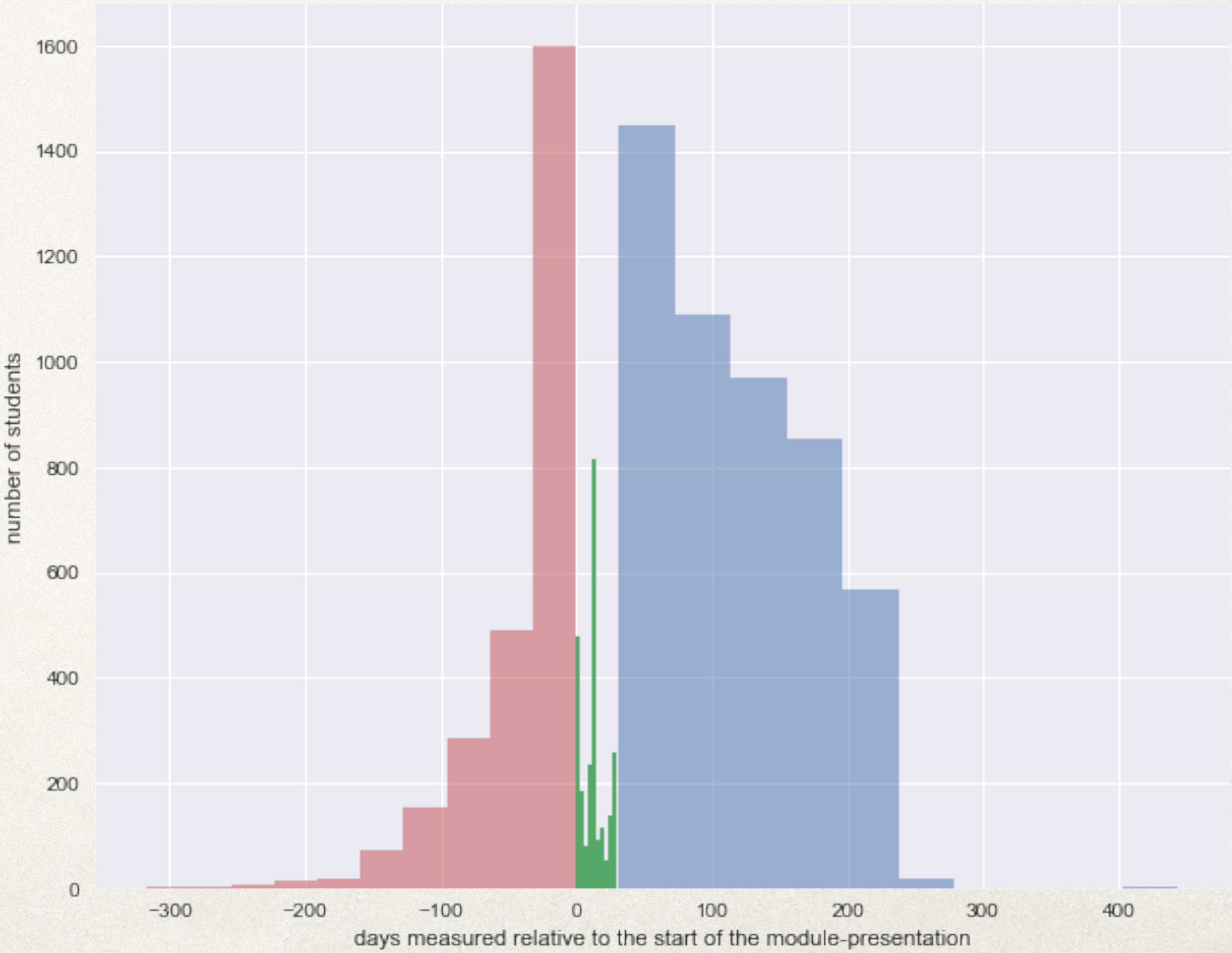
The next data story is done in 5 parts-correlation between student withdrawal and:
* date of withdrawal
* students who are disadvantaged due to their background
* scores and student withdrawal
* VLE engagement and student withdrawal
* disabled students analysis

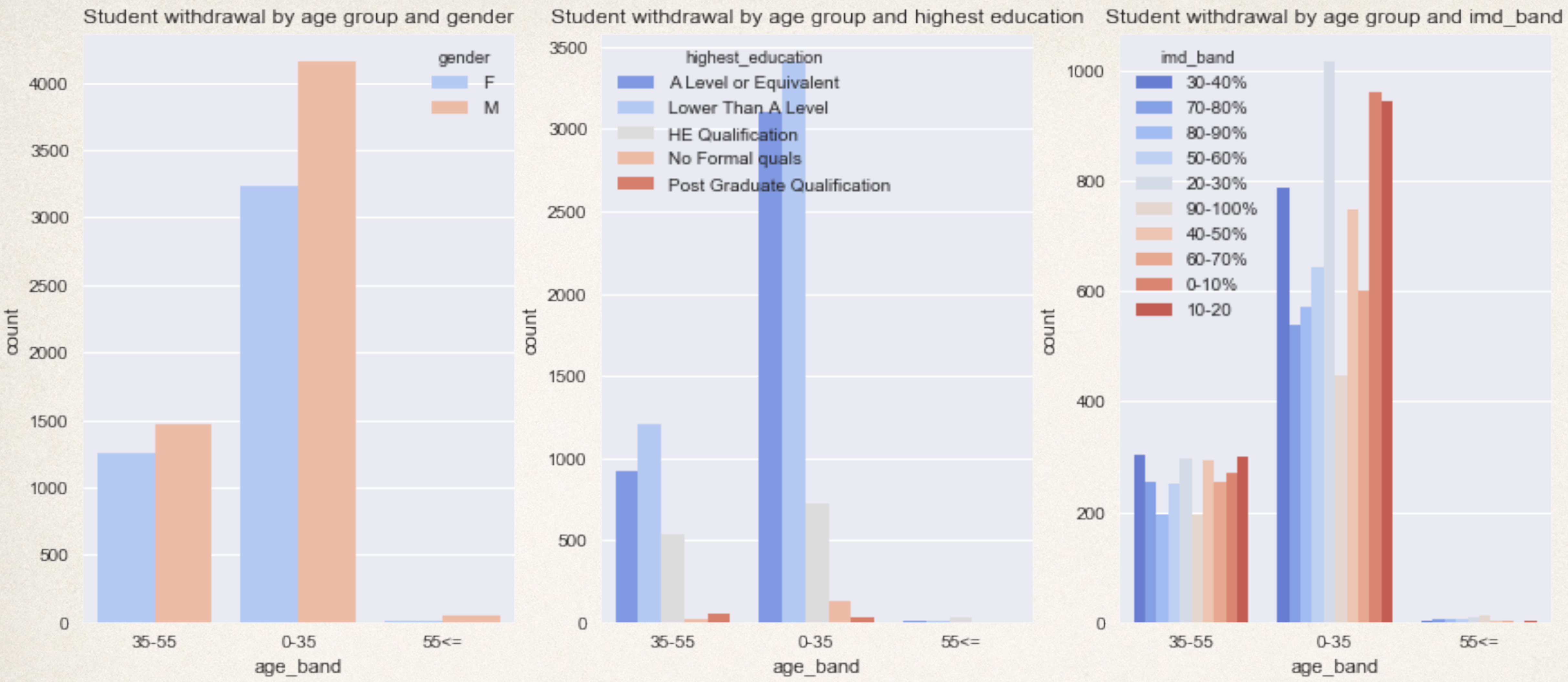
Data is taken from column date_unregistration from studentRegistration.csv: date of student unregistration from the module presentation (this is the number of days measured relative to the start of the module-presentation)

negative number implies students withdrew before the start of module 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	





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

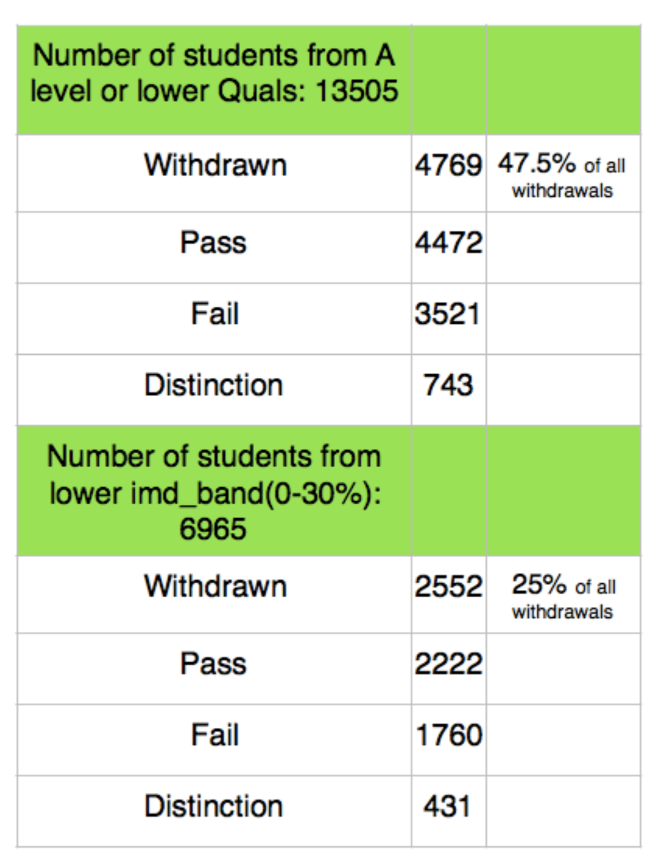
We define these as disadvantaged students or those most at risk.

Out of 32593 enrolled students, 17127 or 52.5% deemed at risk as they have both educational as well as imd_band disadvantage

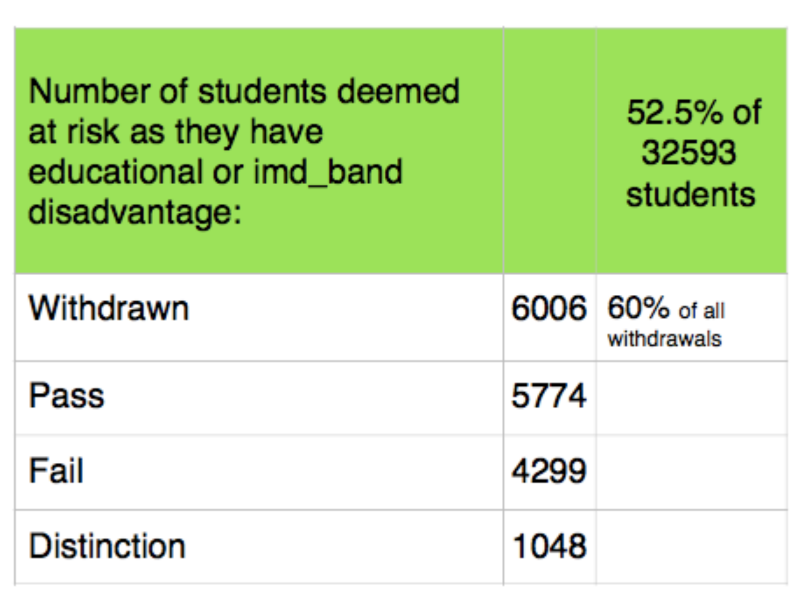
Among these 17127 students, 6006 withdrew or 35%.

Out of 10033 students who withdrew overall, 59% are in the disadvantaged category.
Number of previous attempts is much more for disadvantaged students (after 3)

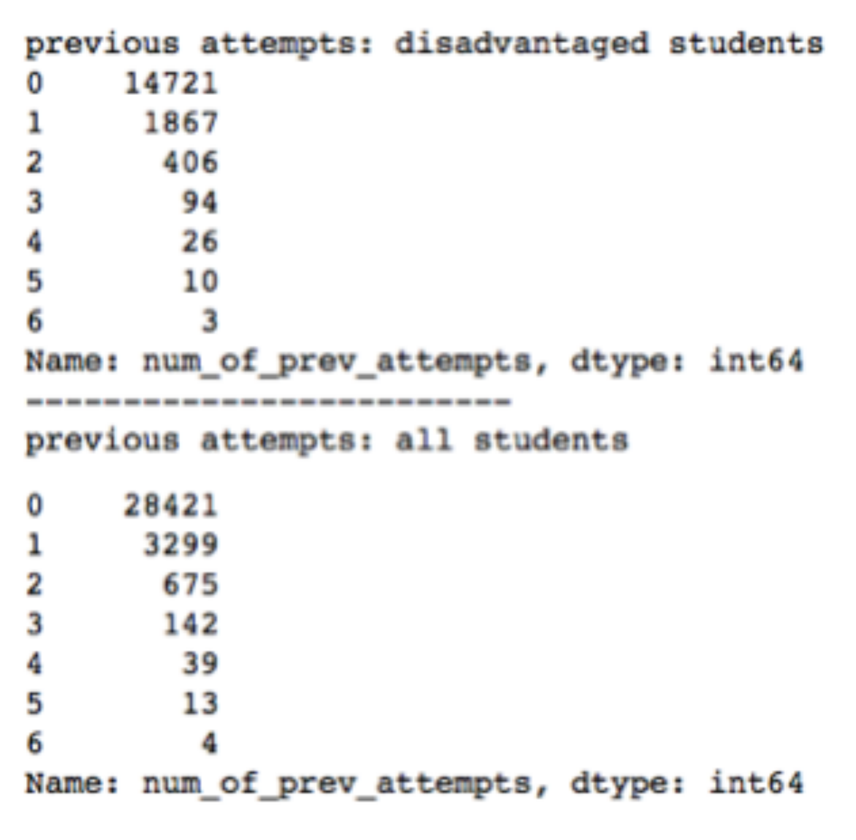
To predict the probability that a student will withdraw,
we will use the above information.



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	

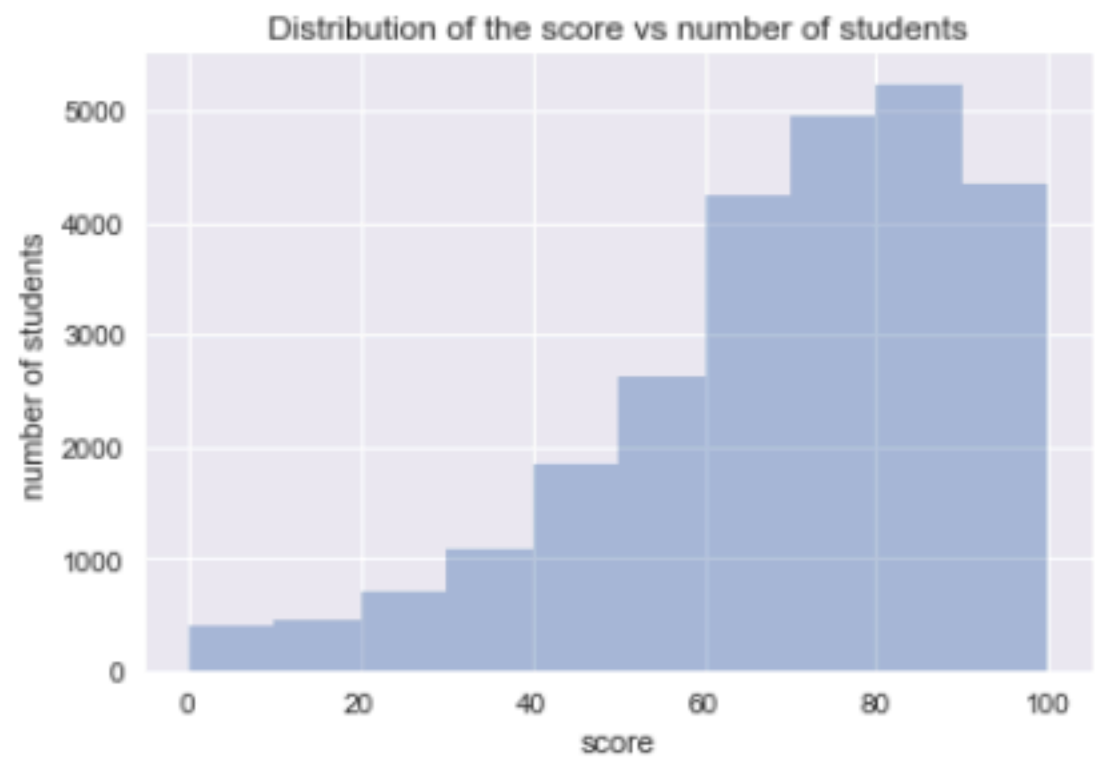


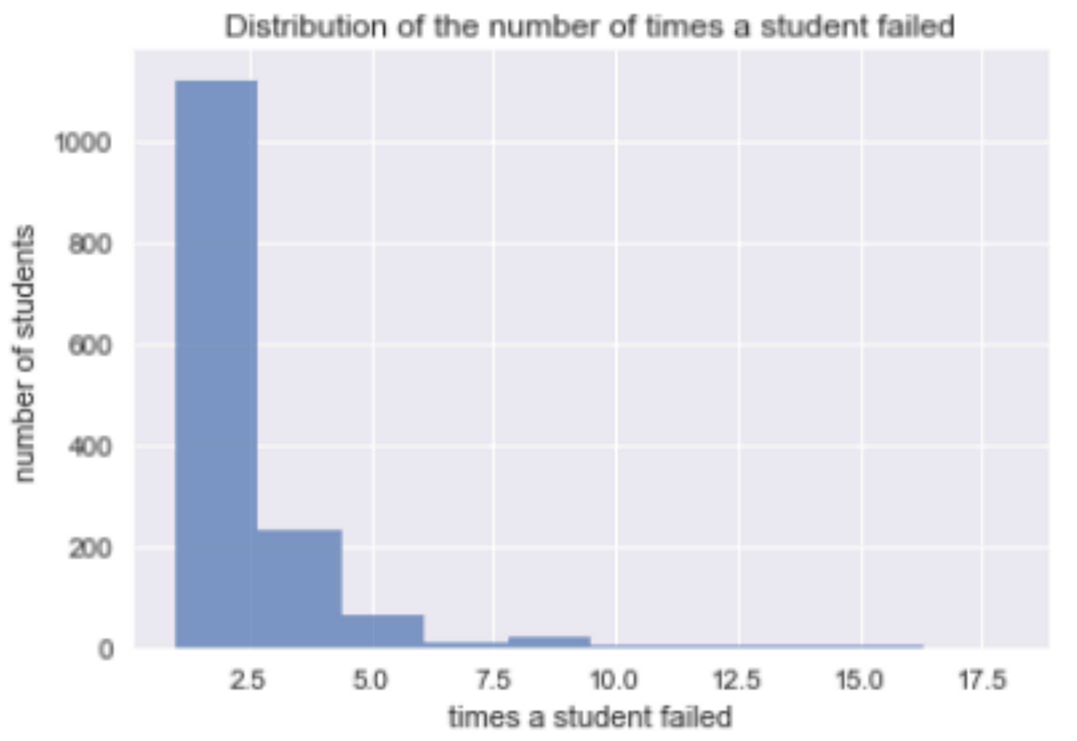
```
previous attempts: disadvantaged students
0    14721
1     1867
2      406
3       94
4       26
5       10
6        3
Name: num_of_prev_attempts, dtype: int64
------------------------
previous attempts: all students

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

Many attempts in the same module can either mean :

- Student failed multiple times: we can see from the graphs that it is not the case. Not many students have failed more than thrice and even then, most students who withdrew did not fail.

- Student took the course but did not complete assessments: only 12.5% of the scores belong to students who withdrew. So this is true. Only 4854 students or about half have a score.



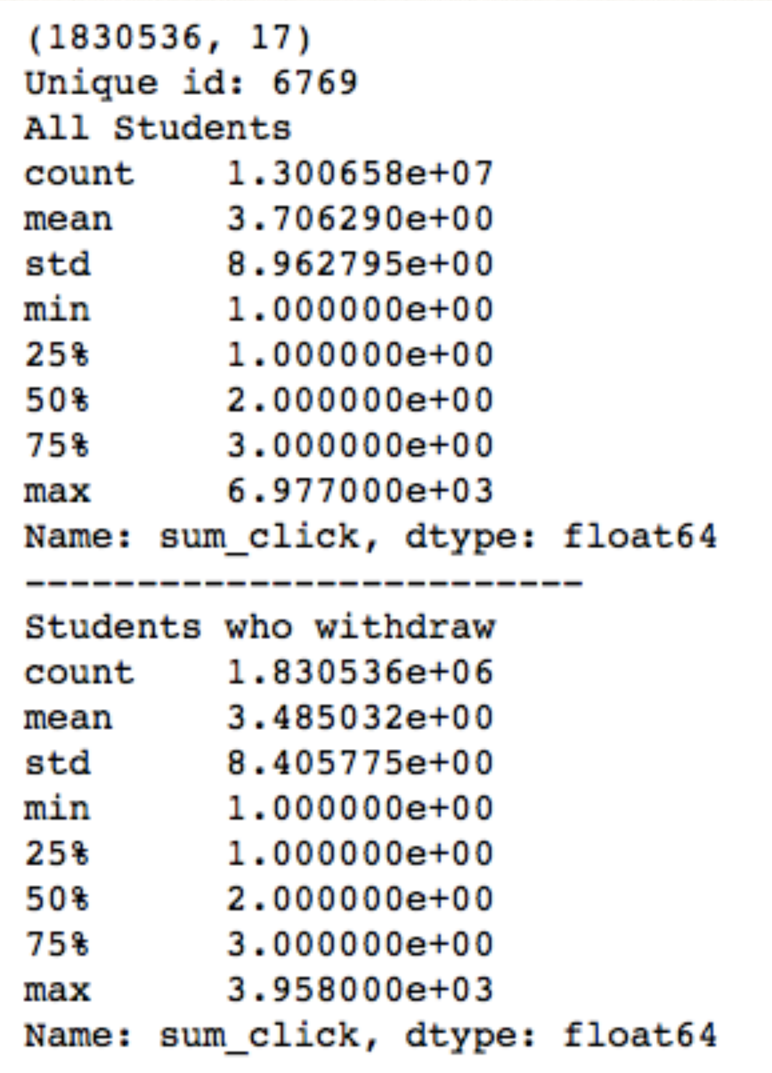


The only VLE engagement data provided is the sum_clicks, total number of clicks by a student.

As we can see from the descriptive stats, the difference between all students vs the ones who withdrew is not vast.

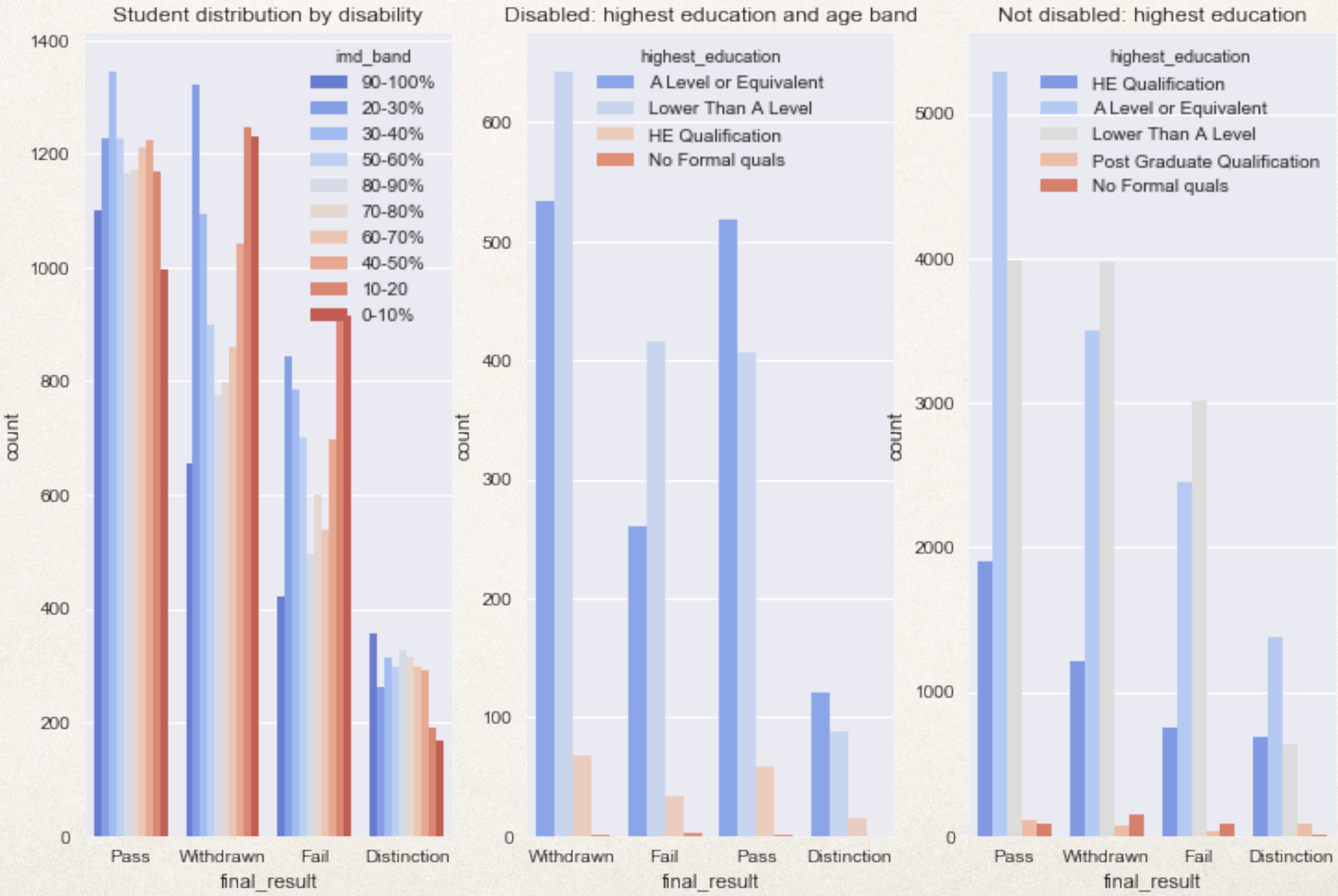
We can be sure that more VLE engagement will have a positive impact on students experience.

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
```

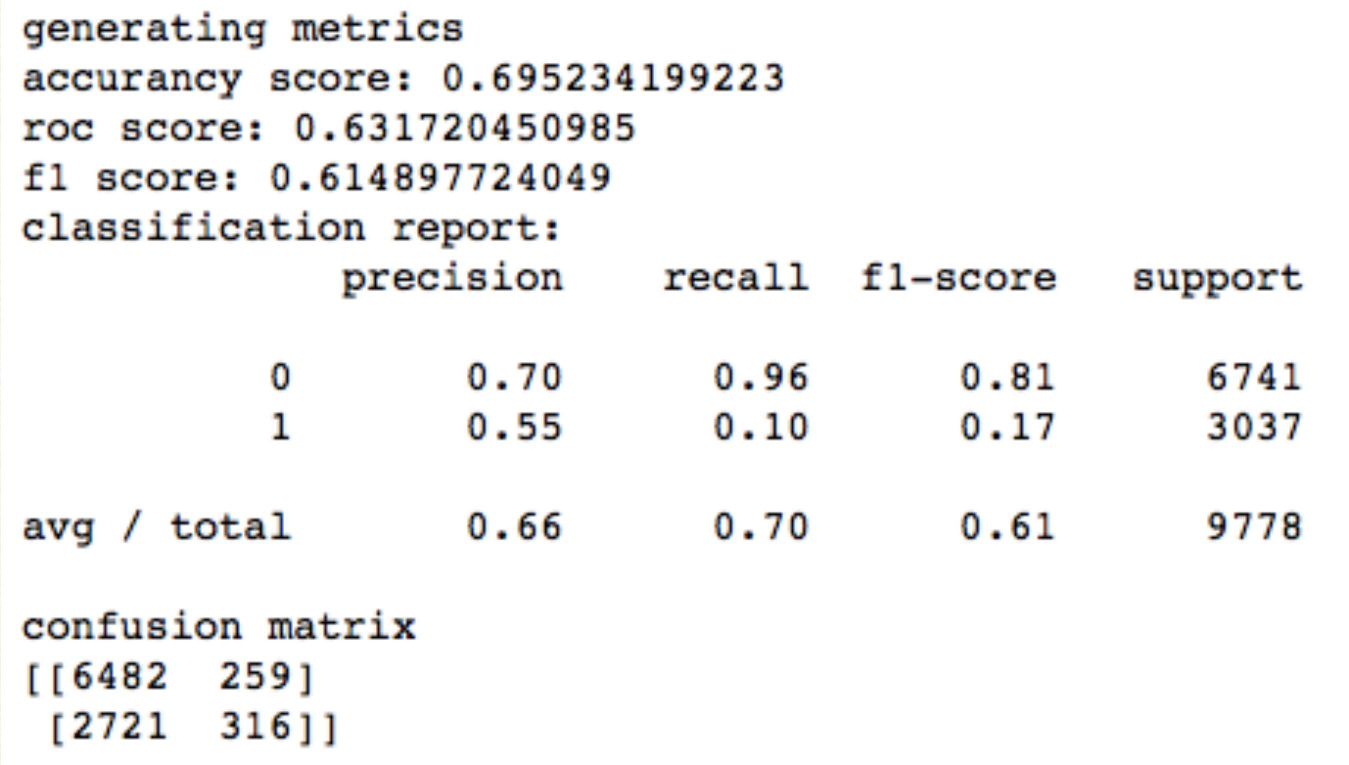
The disadvantaged students with A level or Lower withdraw more than any other group. With more data, it is possible to look at student retention as a separate topic.



Prediction

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.

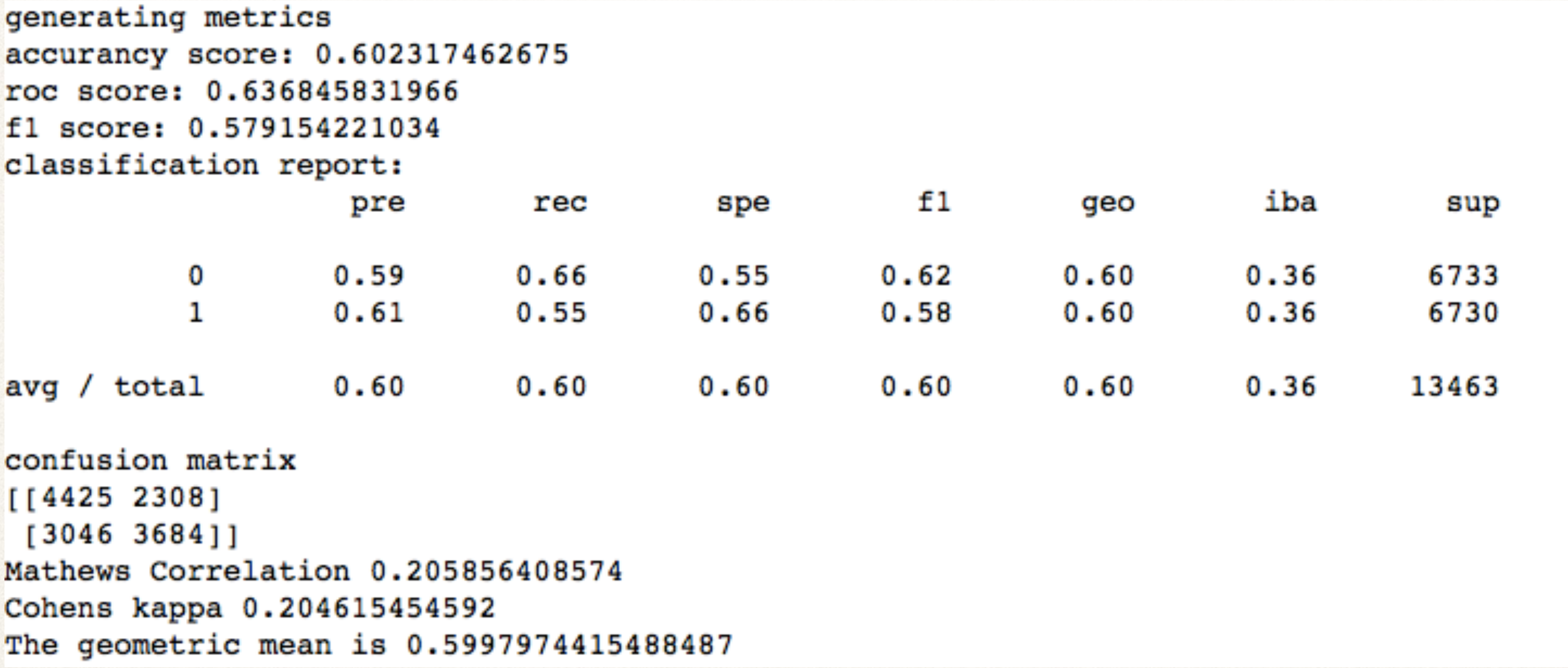
Using Logistic Regression



```
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]]
```

Precision and Recall scores do not match
f1 scores for 0(non-withdrawals) and 1(withdrawals) are disproportionate
Confusion matrix shows the same
Shows data is imbalanced.

Using Logistic Regression using over-sampled data obtain by SMOTE



```
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
```

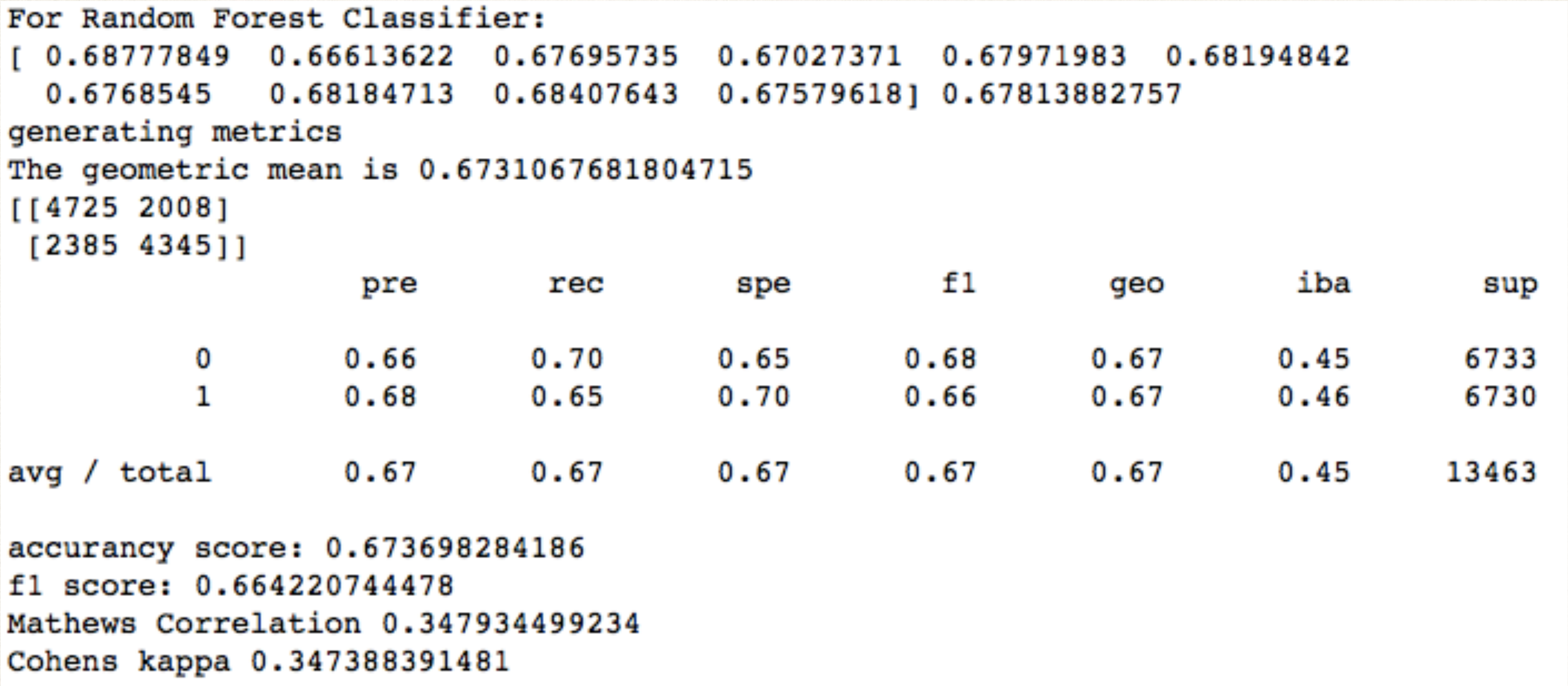
Precision and Recall scores match

f1 scores for 0(non-withdrawals) and 1(withdrawals) are not disproportionate

Confusion matrix shows the same

Can improve MCC, Cohen’s kappa and accuracy

Trying to improve the model using other classification algorithms: 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
```

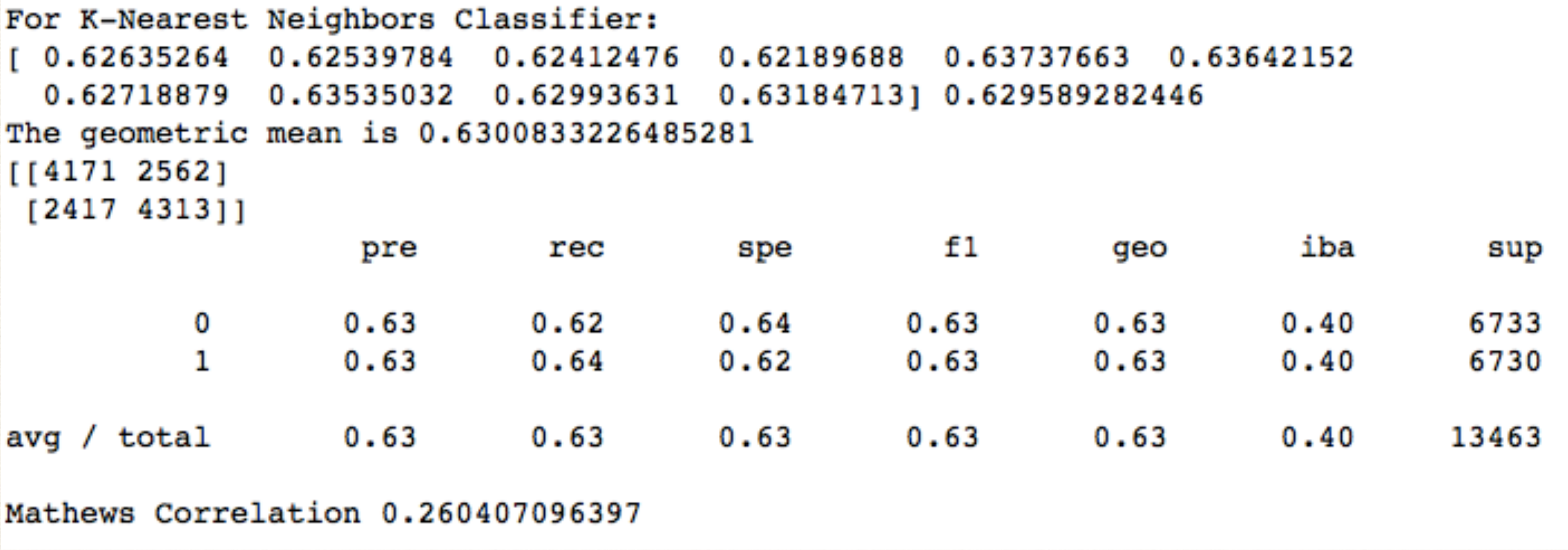
Precision and Recall scores match

f1 scores for 0(non-withdrawals) and 1(withdrawals) are not disproportionate

Confusion matrix shows the same

MCC, Cohen's kappa are fair (0.35)accuracy 67% (better than LR)

Trying to improve the model using other classification algorithms: 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
```

Precision and Recall scores match

f1 scores for 0(non-withdrawals) and 1(withdrawals) are not disproportionate

Confusion matrix shows the same

MCC, Cohen's kappa are fair (0.26)

accuracy 63% (better than LR but not Random Forest)

Recommendations

- # 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 **registration 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: Proactive: 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: Reactive: **Tutorials and peer-peer engagement** to motivate the students and make sure the next two recommendations are followed.
- # 4: Reactive: **Monitor VLE engagemen**t, 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: Proactive: 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 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

- Look at other sampling techniques and classifiers to improve accuracy of prediction

Conclusion

- Student retention in distance learning courses like the Open University, requires a different approach when compared to traditional university structure.

- 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 Logistic Regression to predict with 69% 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