Spring Board Data Science Career Track - Capstone Report

Performance Assessment of Online Training Services Using Data Analytics and Machine Learning Techniques

Jeffrey Ma

Abstract

Many people are enrolled in online courses every day. Schools and specialized organizations are finding a lot of businesses in offering certifications and degrees online. For this capstone project, I wanted to identify factors affecting the performance of online training offers and find ways to improve it. More specifically, I will try to predict withdrawal rates of students enrolled in an online course, so that we can identify students at risk of withdrawing, and intervene in a timely manner to help them to succeed in the course. This information is useful for Massive Open Online Courses and companies that provide online educational services like eDX and Coursera. Identifying students at risk of withdrawing from the course and intervening to help those students succeed can only increase user satisfaction and profit for the training company as more students will enroll in, pass the course, and achieve their goals in life.

Introduction

To catch up with technology development and to enhance their technical competitiveness or enrich their personal life, many people are enrolled in online courses every day. Schools and specialized organizations are finding a lot of businesses in offering certifications and degrees online. However, a persistence challenge for both the students and service providers are that the dropout rate has been persistently high (Willging and Johnson, 2005; Qiu et al., 2019).

The main objectives of this project are to identify the main factors affecting students' dropout rate so that online training course providers may find the best ways to intervene before a student is dropped out. Since this project is part of the Springboard training program, another objective of this project and the report is to demonstrate in a systematic way what was learned in the program.

Dataset

To accomplish such a task, we need a proper dataset. The one used for this project is the anonymized Open University Learning Analytics Dataset (OULAD). It contains data about courses, students and their interactions with Virtual Learning Environment (VLE) for seven selected courses (called modules). Presentations of courses start in February and October; marked by "B" and "J" respectively. The dataset consists of tables connected using unique identifiers. All tables are stored in the csv format.

Data Wrangling

The data wrangling portion of the capstone project had two main objectives. The first objective was to remove extraneous data. Columns containing redundant data or non-relevant descriptive information like student ID's were removed. The second objective was feature generation. Using a label encoder, ordinal variables were mapped to integers. Also nominal values were encoded as dummy variables through one hot encoding. The target column of final result of withdraw or not withdraw was prepared. Finally the number of clicks students made in the course were summed and used as a feature.

After preprocessing, a data frame was used containing columns with information on

1. age,
2. gender,
3. region,
4. level of education,
5. disability,
6. number of previous attempts,
7. sum clicks, and
8. final result

One problem during data wrangling that was resolved was that there were duplicate rows due to a student being enrolled in more than one course. To fix this issue, a dummy variable was created where we coded "passing" and "passing with distinction" as zero, and "failing" and "withdrawing" as a 1. Using this dummy variable, we could select the max value of 1's of the people who withdrew or failed any course. Thus giving us a table where each row indicates the performance of the student in the course for a module. Below Table 1 is a look at our dataframe with dummy variables included.

Column 1: ID of Student

Column 2: Studied Credits of Student

Column 3: Final Result after using label encoder to convert Pass/Distinction/Withdraw/Fail into 1/2/3/4

Column 4: sum of the number of clicks each student made

Column 5: Feature created to convert passing and passing with distinction into a zero, and Withdrawing and failing into a one.

Column 6: After converting region to scalars with a label encoder, used one hot encoding to get more dummy variables

Column 7: more dummy variables, etc.

Table 1 Example of raw data

```
combined.head()
```

id_student	studied_credits	final_result	sum_click	final	region_0	region_1	region_2	...	age_band_0	age_band_1	age_
59	4	3	885	1	1	0	0	...	1	0	0
59	4	1	885	0	1	0	0	...	1	0	0
102	4	3	2236	1	0	0	0	...	0	1	0
102	4	2	2236	0	0	0	0	...	0	1	0
116	4	3	880	1	0	0	0	...	0	1	0

Exploratory Data Analysis

Look at Final Result, target variable of interest

Final Result		Percentage of Total
Withdrawn	16697024	62%
Pass	6924889	25.7%
Fail	2169591	8%
Distinction	1149815	4.3%

```
#select dataframes to combine

combined = [student_assessment, student_info, student_registration, student_vle]
combined = reduce(lambda left,right: pd.merge(left,right,on='id_student'), combined)
combined['final_result'].value_counts()
```

```
Withdrawn      16697024
Pass            6924889
Fail            2169591
Distinction     1149815
Name: final_result, dtype: int64
```

In total, approximately 70% of the students failed to pass the course. Only approximately 30% of all students enrolled in the course finished with a passing grade.

Course Enrolled. During exploratory analysis of our dataset, bar graphs comparing final results in the course with various independent variables such as gender, age, and region were produced. Figure 1 is an example.

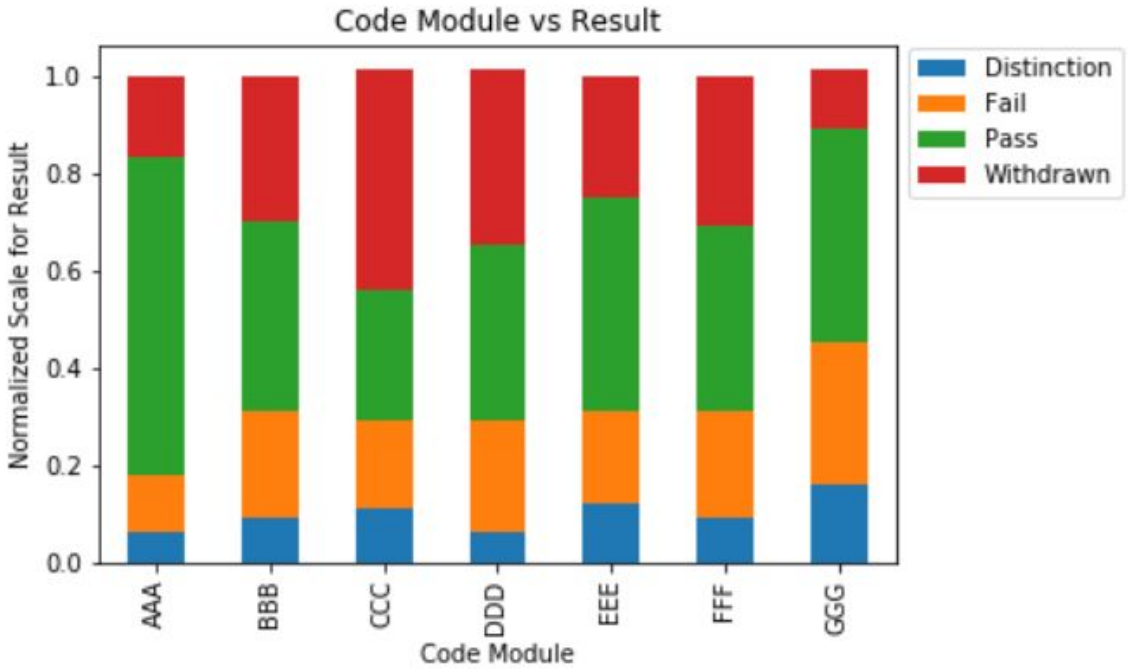


Figure 1 Normalized Bar graph shows student performance in each of the enrolled seven courses. Color blue indicates passing with distinction, orange fail, green pass, and red withdraw.

From Figure 1, we can see that the distribution of students who withdrew (red) and passed green) is roughly the same for all the modules or course.

Gender. Figure 2 shows statistical performance of male and female students enrolled in online training courses. From this figure, it appears that, understandably, gender has little effect on the performance of students.

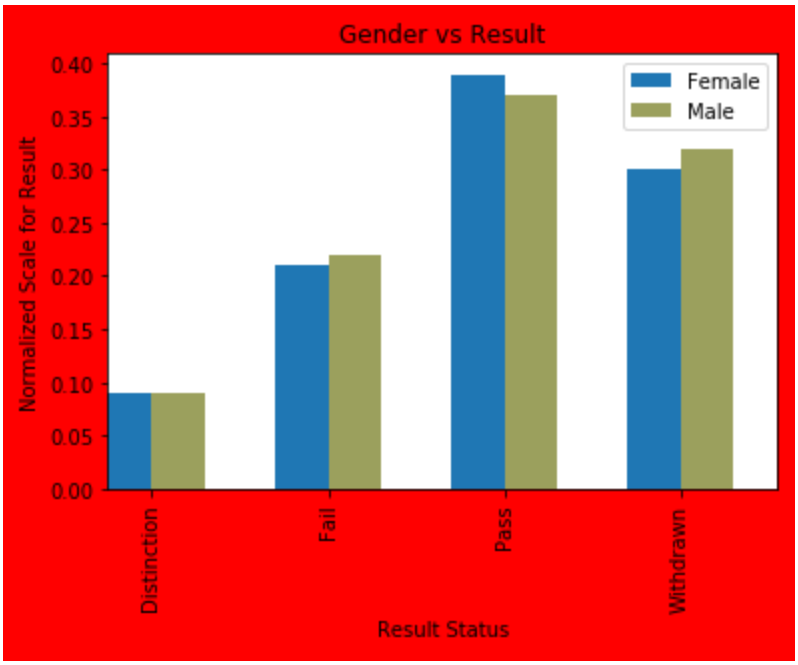


Figure 2 Bar graph shows performance of male and female students. Color blue indicates passing with distinction, orange fail, green pass, and red withdraw.

Region. Figure 3 shows the performance of students from different regions. It looks roughly even among the regions for distribution of pass rates as well. No single group looks distinct.

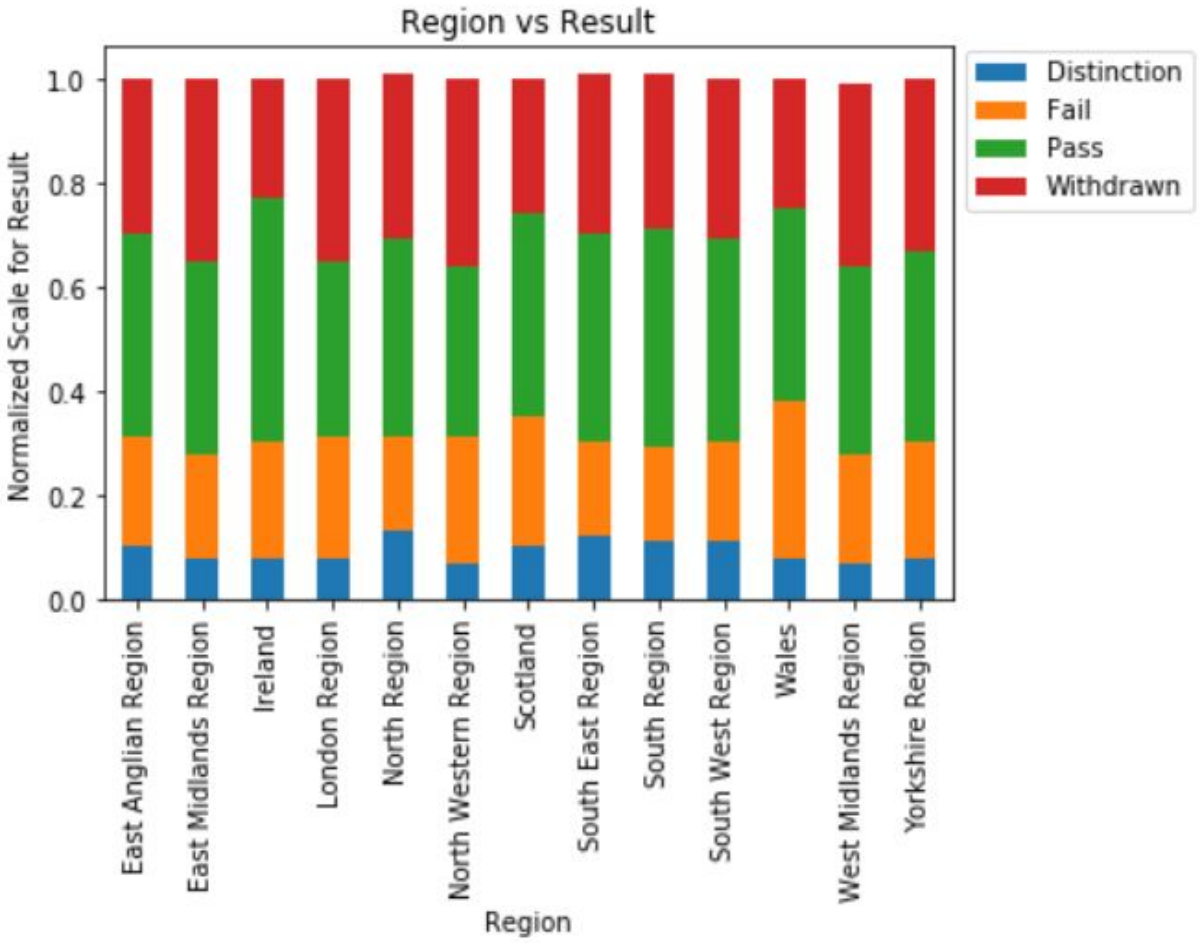


Figure 3 Bar graph shows performance of students from different regions. Color blue indicates passing with distinction, orange fail, green pass, and red withdraw.

Educational Background.

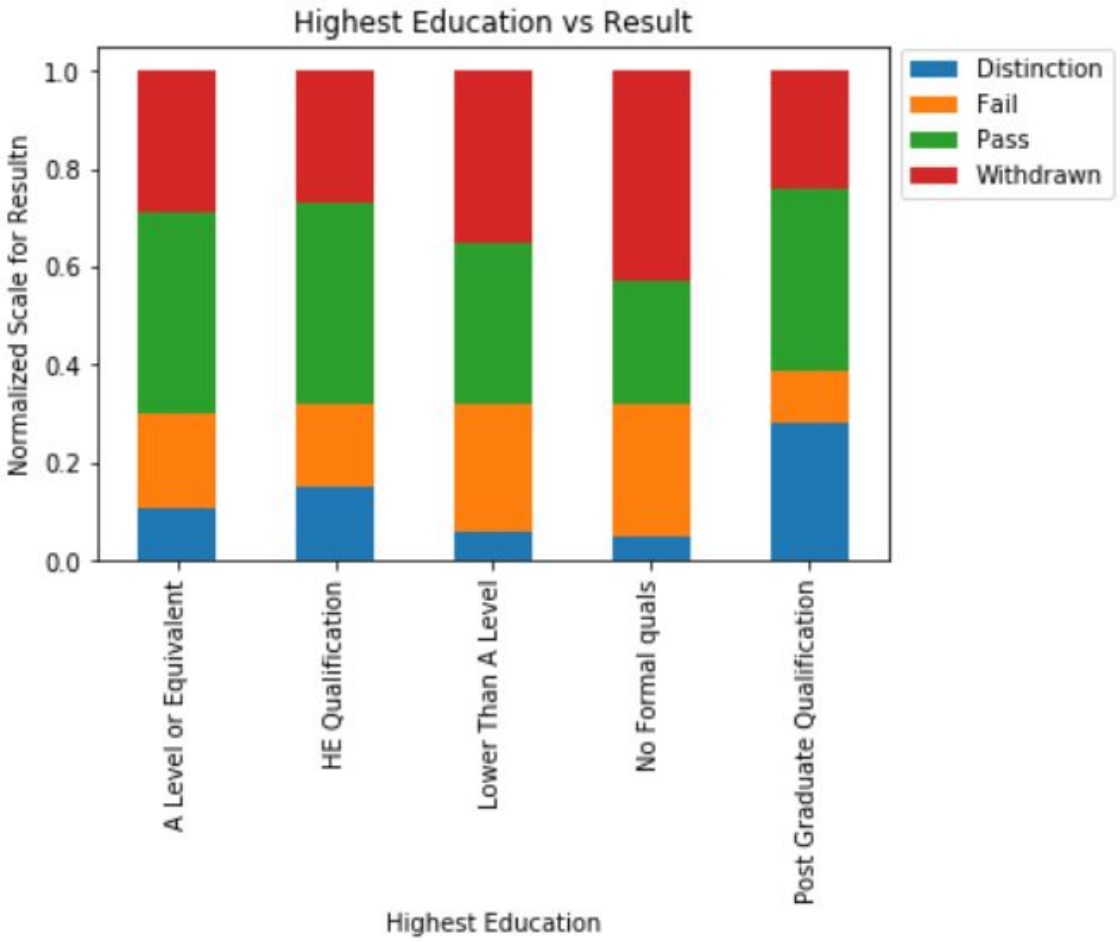


Figure 4 Bar graph shows performance of students from different educational backgrounds. A Level or Equivalent represents ..., HE Qualification ..., Lower Than A Level ..., No Formal qual, and Post Graduate Qualification ... Color blue indicates passing with distinction, orange fail, green pass, and red withdraw.

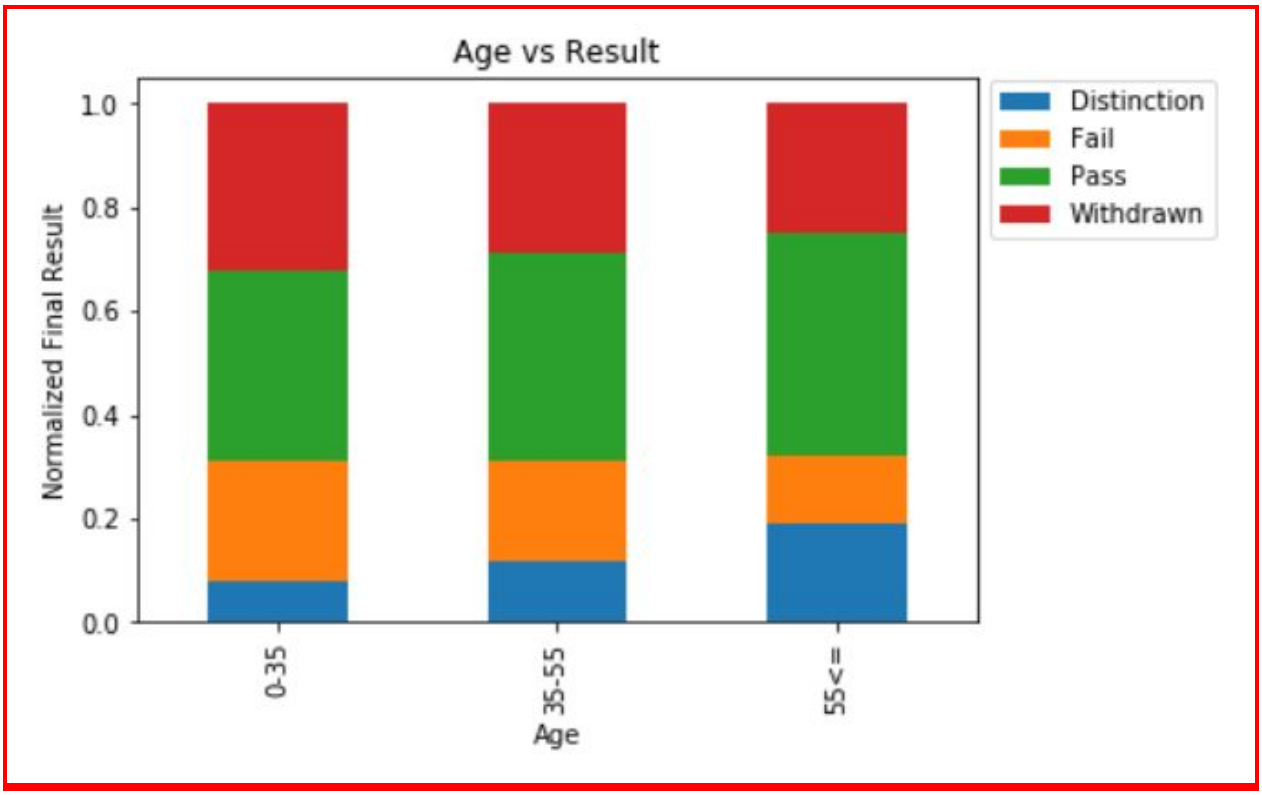
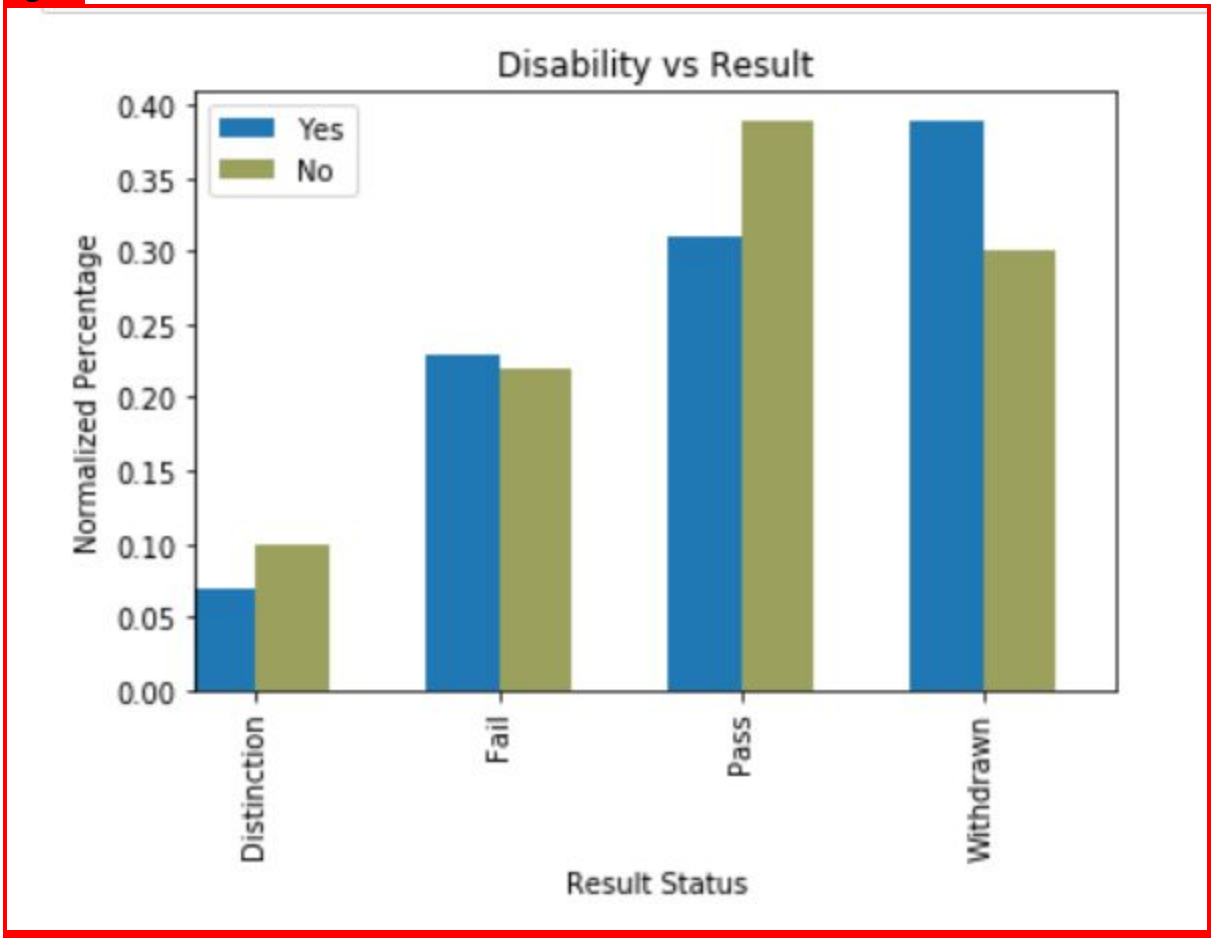


Figure 5 Bar graph showing performance of students of different age brackets. It can be seen the distribution of passing and failing is roughly even

Figure



Variable Correlations. The various correlations between the independent variables can be seen from Table **2**.From Table 2 there appears to be a correlation between the number of clicks and the student's final performance in the course.

Table 2 Correlations between different independent variables

	id_student	highest_education	imd_band	age_band
id_student	1.000000	0.042191	0.024396	0.198493
highest_education	0.042191	1.000000	0.125717	0.110026
imd_band	0.024396	0.125717	1.000000	0.071579
age_band	0.198493	0.110026	0.071579	1.000000
num_of_prev_attempts	0.012470	-0.041709	-0.040758	0.005033
studied_credits	-0.006083	0.010024	-0.037872	-0.075143
Distinction	0.028164	0.127624	0.086055	0.064893
Fail	-0.032051	-0.098357	-0.082725	-0.053291
Pass	0.003417	0.062797	0.070033	0.029214
Withdrawn	0.007341	-0.064971	-0.059340	-0.026833
sum_click	0.037327	0.082098	0.077465	0.140429

	num_of_prev_attempts	studied_credits	Distinction
id_student	0.012470	-0.006083	0.028164
highest_education	-0.041709	0.010024	0.127624
imd_band	-0.040758	-0.037872	0.086055
age_band	0.005033	-0.075143	0.064893
num_of_prev_attempts	1.000000	0.181726	-0.068826
studied_credits	0.181726	1.000000	-0.057721
Distinction	-0.068826	-0.057721	1.000000
Fail	0.094794	-0.029222	-0.181751
Pass	-0.073689	-0.092765	-0.283667
Withdrawn	0.039994	0.172529	-0.197805
sum_click	-0.068865	-0.006473	0.257051

	Fail	Pass	Withdrawn	sum_click
id_student	-0.032051	0.003417	0.007341	0.037327
highest_education	-0.098357	0.062797	-0.064971	0.082098
imd_band	-0.082725	0.070033	-0.059340	0.077465
age_band	-0.053291	0.029214	-0.026833	0.140429
num_of_prev_attempts	0.094794	-0.073689	0.039994	-0.068865
studied_credits	-0.029222	-0.092765	0.172529	-0.006473
Distinction	-0.181751	-0.283667	-0.197805	0.257051
Fail	1.000000	-0.455504	-0.317630	-0.209286
Pass	-0.455504	1.000000	-0.495738	0.280061
Withdrawn	-0.317630	-0.495738	1.000000	-0.299305
sum_click	-0.209286	0.280061	-0.299305	1.000000

Summary of exploratory findings:

- 62% of the students withdrew from the course and 8% failed out of the course. This means only approximately 30% of the students registering for the course made it through to the end. Hopefully through our intervention in the future, we can increase these numbers.

- From Table 2 there appears to be a correlation between the number of clicks and the student's final performance in the course.
- Students are distributed relatively evenly over different geographic regions (Figure 3).
- Men and women appear to have similar pass and fail rates
- Most of the people taking the course have A level (secondary school level) education or lower (Figure 4) indicating these are technical training courses.

After exploratory analysis, we were comfortable selecting various independent variables from the combined dataset and building a machine learning model off them to predict withdrawal rates of students from the course.

Predictive Modeling with Machine Learning

Data Preparation. After preprocessing of the data in the data-wrangling portion of the project, a data frame where each row represented one student's performance in one module of the online course was left for machine learning modeling.

Model Building and Validating. The data was split into a training and a test set, and then a decision tree classifier and random forest classifier were run on the training set. The predictive model for student's withdrawal rates used the independent variables of

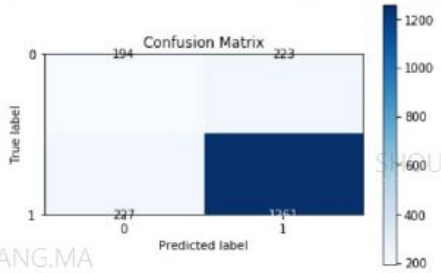
1. gender,
2. region,
3. highest education,
4. age band,
5. number of previous attempts,
6. studied credits, and
7. disability.

After training a decision tree classifier as well as a random forest classifier on the train set, we evaluated the precision of the models built. When comparing precision of the Random Forest Model in Figure 5 with the Decision Tree Model in Figure 6, it was found that the random forest classifier model had the highest cross validated precision of 0.85.

```
y_pred = cross_val_predict(rf, xTrain, yTrain)
skplt.metrics.plot_confusion_matrix(yTrain, y_pred, normalize=False)

plt.show()
```

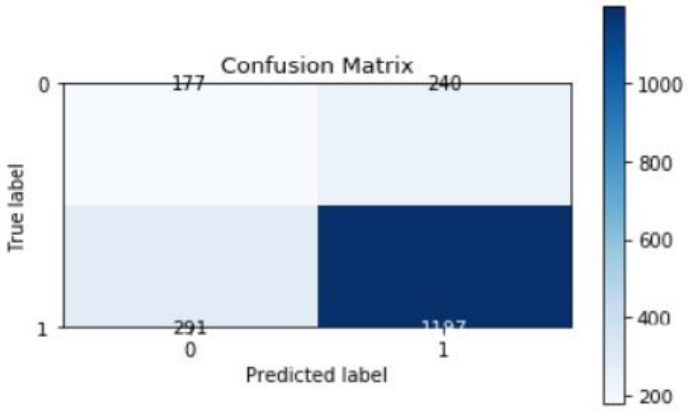
```
C:\ProgramData\Anaconda3\lib\site-packages\sklearn\model_selection\_split.py:1978: FutureWarning: The default value of cv will
change from 3 to 5 in version 0.22. Specify it explicitly to silence this warning.
  warnings.warn(CV_WARNING, FutureWarning)
```




```
print("Precision:{0:.3f}".format(metrics.precision_score(yTrain, y_pred)),"\n")
```

```
Precision:0.850
```

Figure 5 Cross Validation of Random Forest Model with a Precision for Prediction of 0.850.




```
2]: print("Precision:{0:.3f}".format(metrics.precision_score(yTrain, y_pred)),"\n")
```

```
Precision:0.833
```

Figure 6 Cross Validation of Decision Tree Model with a Precision for Prediction of 0.833.

Consequently, the random forest classifier models were used for prediction.

Model Enhancement. A second set of models were built using the same independent variables as before except we also added the sum_clicks variable to see if more information and a better prediction could be obtained. After running the model on the test data set, we achieved a slightly improved precision of 0.852 (Figure 7) on predicting withdrawal rates of students in online courses, demonstrating that the model can be used to confidently predict withdrawal rates from registered online training courses.

```
In [219]: # Random Forest for Table with sum_click
          X = combined.loc[:, combined.columns != 'final']
          y = combined['final']
          xTrain, xTest, yTrain, yTest = train_test_split(X, y,train_size = 0.75)

          rf = RandomForestClassifier(n_estimators=10,random_state=33)
          rf = rf.fit(xTrain, yTrain)
          train_pred = rf.predict(xTrain)
          test_pred = rf.predict(xTest)
          print("Precision:{0:.3f}".format(metrics.precision_score(yTest, test_pred)),"\n")

          Precision:0.852
```

Figure 7 ....

Conclusion and Recommendations

Based on this data analytics study, the following are concluded,

- The successful rate of a student enrolled in an online training course and receive a passing grade is about 30%, with about 62% withdraw before completing and about 8% fail the course requirement.
- Gender and region where the student is located do not seem to play a role in the above statistics.
- Students who seek help measured by clicks clocked by the online training website tend to be more successful in completing the enrolled courses, providing a hint to improve the successful graduation rate.
- With the machine learning model built, students at risk of withdrawing from the course were identified and intervention could then be conducted to help these students graduate from the course. This is useful for any online teaching company that wishes to improve their user's experience and success, thus their own business success.
- Intervention could be technical help, spiritual encouragement could equally be effective.
- Engagement of any kind from the student with course material increased that student's probability of passing the course. This information can also be used for students who wish to successfully pass the online training program of their choice.
- In the future, more experiments could be conducted to find other independent variables that have a positive effect on student's matriculation rates. Some example features that come to mind are pre-online-training educational grade point average and income bracket.

Acknowledgement

I would like to express my deep appreciation to my mentor Mr. Nate Sutton for his guidance not only during this project, but across the whole Springboard program. Thanks also to my Springboard counselors and program managers for their assistance. Special thanks goes to my family for their encouragement on this journey to be a data scientist.

References

Qiu, L., Liu, Y., Hu, Q., and Liu, Y., Student dropout prediction in massive open online courses by convolutional neural networks, in Soft Computing, v.23, issue 20, Oct 2019, pp.10287-10301.

Willging, P.A. and Johnson, S.D., Factors that influence students' decision to dropout of online courses, Journal of Asynchronous Learning Networks, v.3, Issue 3, pp.115-127, 2005.