Identification of Key Drivers of Student Success

Prepared By:
Shiv Prakash Ojha

Executive Summary

An extensive analysis of the Open University Learning Analytics Dataset (OULAD) was performed using data exploration, machine learning and causal analysis to identify the drivers of student success. The best model using gradient boosting method showcased an accuracy of around 90 percent for prediction of successful completion of course by a student.

Results show that a high interaction with the Virtual Learning Environment (VLE) is a good indicator of success for students. Higher values of Index of Multiple Deprivation (IMD) band, higher education levels, and absence of disability are recognised as key drivers of success for students by predictive modelling and causal analysis. Region, age, and gender have a comparatively lower impact on the chances of success. Based on these observations, it is recommended to invest in opportunities to assist students with a background of higher deprivation or those having disability. The results also show impact of subjects on student performance which indicate a need for uniform marking policies for all subjects. Usage of the model for investigating efficacy of opportunities for improving student performance is a recommended future project.

The analysis is limited in its scope by the data used. Important aspects of education such as non-digital forms of education like availability of reference books and their impact on results are not considered. In addition to highest education level, relevance of previous education to current course also needs consideration. Evaluation of student aptitude in the form of previous assessment scores may be useful additions to the dataset. These will help to further improve the analysis results by identifying other drivers of student success like student aptitude or efficiency of different modes of education for different students.

Analysis

1. Data Exploration and Pre-Processing

The data consists of tables with information on students, courses, assessments, and Virtual Learning Environment (VLE) including around 32,600 student records of around 28,800 unique students. Socio-economic background of students is included in the form of their gender, age, highest education level, previous credits completed, region of living, Index of Multiple Deprivation (IMD) band, and disability status. For each course enrolled by a student, outcome is defined in categories of Pass, Fail, Withdrawal from course, or achievement of distinction. We assume definition of success for a student as the successful completion of course (“final_result” of “Pass” or “Distinction”).

Preliminary exploration helps to identify potential drivers affecting the success of students. Stacked bar charts showing percentage of students having different outcomes of the courses from each category are used to compare the impact of each factor on success of students. Figure 1 shows some of the factors with impact readily visible on student success. IMD Band and disability are seen to have a clear impact on the success of student. Age and highest education levels also showed indications of impact on student performance.

The potential drivers of impact are identified using preliminary data exploration, but a quantitative estimate cannot be obtained. The effect seen in plots may be a combined effect of multiple factors. Hence, modelling is used to estimate the impact of individual factors on the success of students.

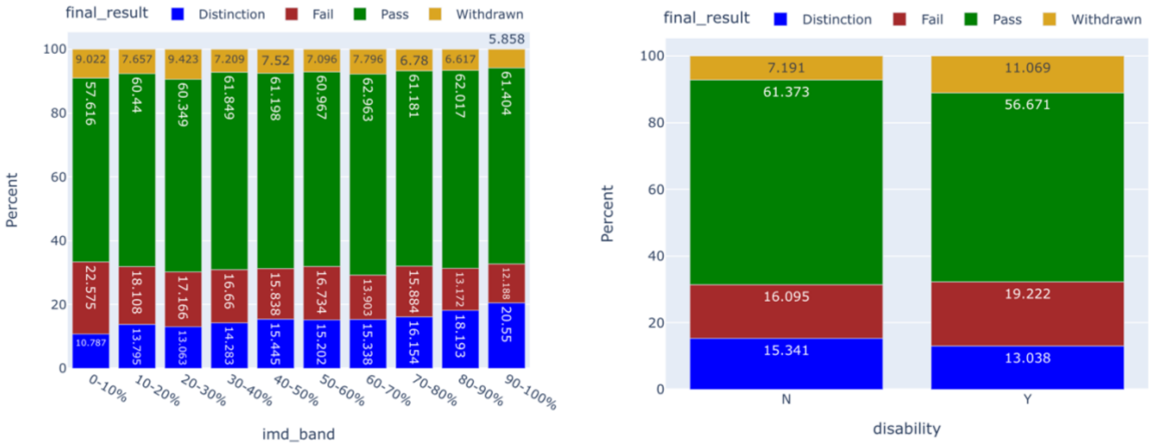

Figure 1: Impact of a) imd_band, and b) disability on the outcome of students belonging to each group. Increase in IMD band shows better result from students. Disability is seen to have a detrimental impact on student performance.

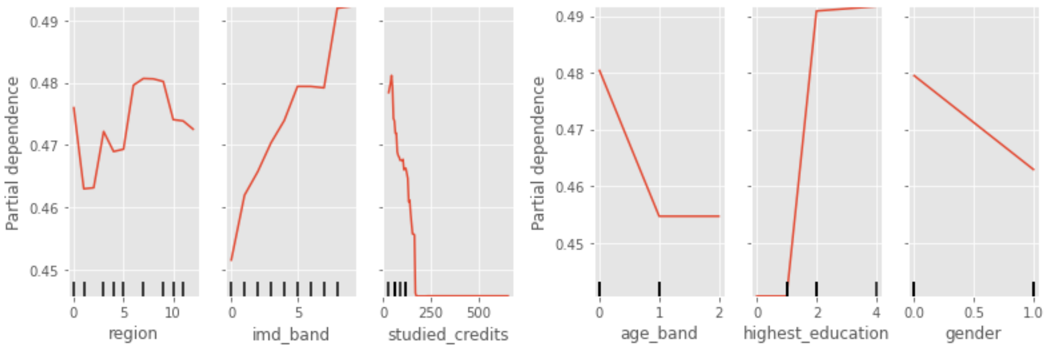

Figure 2: Partial Dependence plots showing relationship of individual attributes with chances of success for students.

2. Predictive Modelling

Predictive models including linear probability models, and machine learning techniques like support vector machines, random forest, and gradient boosting methods were used. The objective here was to model the complex relationships between attributes to identify most important drivers of success of students and to quantify their relative impact. The data was pre-processed to make it conducive to classification models. Missing values from IMD band were replaced with most frequent value in living region of student. Student data was combined with student interaction data with VLE to create the consolidated dataset for further analysis. Dataset was split into training and test datasets with 80:20 ratio randomly.

Out of all methods used, Light Gradient Boosting Machine (LGBM) showed best results for test data with an accuracy of around 90 per cent. Partial Dependence Plots (PDP) were generated to assess the importance and relative impact of attributes on results of students. PDP plots (Figure 2) showcase the change in probability of success for students with change in values of attributes. Apart from the VLE interaction attributes, region, IMD band, and studied credits as some of the most important drivers of success. Additionally, age, highest education, and gender, are also seen having an impact on the results of the students. However, causal analysis is recommended to differentiate between causation and correlation of attributes.

4. Causal Analysis

Inherent aptitude of students is recognised as a key driver of success (Figure 3). However, the aptitude of students is unobserved in our dataset. The aptitude will manifest in the form of interaction of students with VLE which helps in getting a better outcome. VLE interactions may exhibit higher correlation with the result. However, the mere fact that a student interacts more with VLE does not guarantee a better result which means that VLE interactions do not have a causal relationship with the outcome. Attributes like highest education level, number of previous attempts, and credits studied will impact both the aptitude and outcome and were included in the causal model for calculation of treatment effect. Similarly, attributes like region, IMD band, age, disability, and gender do have an impact directly on the outcome. Module and presentation effects need to be included in the model to estimate marginal effect of other attributes. There may be other unobserved attributes impacting the outcome. However, as the student attributes are likely to be independent of these unobserved attributes, the model results are likely to be reliable.

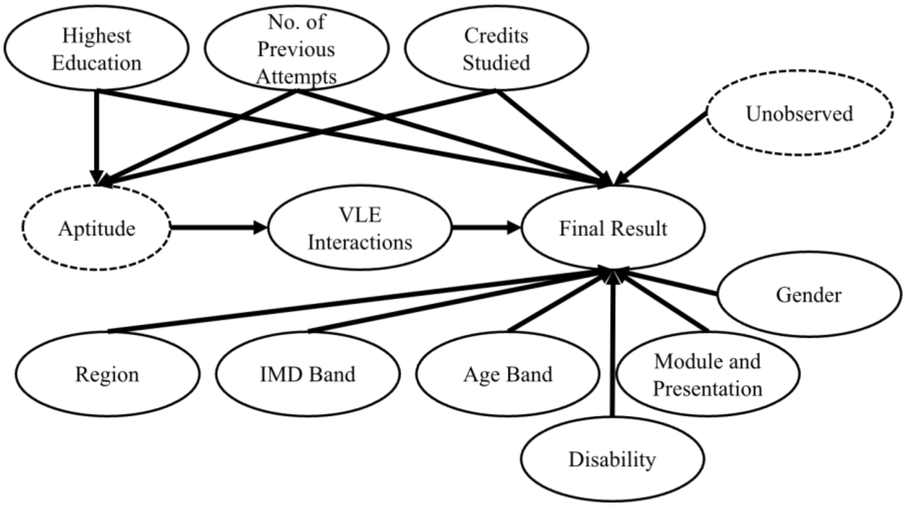

Figure 3: Causal diagram used to differentiate between causation and correlation between different drivers of student success.

A causal forest model is created to estimate effects of the attributes considering the causal relationship shown in Figure 3. VLE interactions are considered as covariates with remaining attributes taken as treatment variables. Comparison of average treatment effect shows that modules contribute to chances of success which highlights need to uniform marking policies across subjects. Highest education levels, IMD band, and disability are identified as most important factors affecting chances of student success. Gender and age are seen to have a smaller impact on student performance. Hence, investment in opportunities to assist students with disability or from areas with higher deprivation index is recommended.

Conclusion

Analysis of the OULAD dataset was performed to identify key drivers of student success. The analysis was performed using data exploration, predictive machine learning models and causal analysis. Analysis indicated that the socio-economic background of students affected their course outcome. Additionally, disability was found to be detrimental to student performance. The analysis also points out that higher interactions with VLE serve as early indicator of success. Based on the analysis, it is recommended to investigate course design and accessibility features to assist students with disability. Further investigation is recommended as to why students from different regions and IMD backgrounds perform with different results and what can be done to facilitate their educational experience. Supplementing the dataset with additional information on availability of alternate forms of education, or student aptitude will also be helpful in improving quality of recommendations.