Student Success Analysis in Online Learning Environment

Tevfik Çağrı Dural

Springboard Data Science Career Track

Branko Kovac

March 2021

Abstract

As the online learning environment was building up since the 2010s with the break of the pandemic in 2020 this field has blown and is expected to rise. As of 2026, it's expected to be $50 billion. In these circumstances, it is critical to understand the success of participants.

This work analyses Open University's Learning Analytic's Dataset[1]with various machine learning models to understand what is important to achieve success in an online learning environment and evaluates the outcomes.

Keywords: online learning, student success, machine learning, data analysis

[1]https://analyse.kmi.open.ac.uk/open_dataset

Problem Statement

At the urge of online learning age how all participants; students, lecturers and service providers understand the student behaviour to reinforce students' success rate? For the data specific approach is to understand all features effect and and relations with the *final_result* feature.

Dataset

The data is Open University Learning Analytics Dataset. Kuzilek, J., Hlosta, M., & Zdrahal, Z. https://doi.org/10.6084/m9.figshare.5081998.v1 (2017). It consists of seven courses for two semesters of 2013 and 2014. Raw data is a set of seven tables.

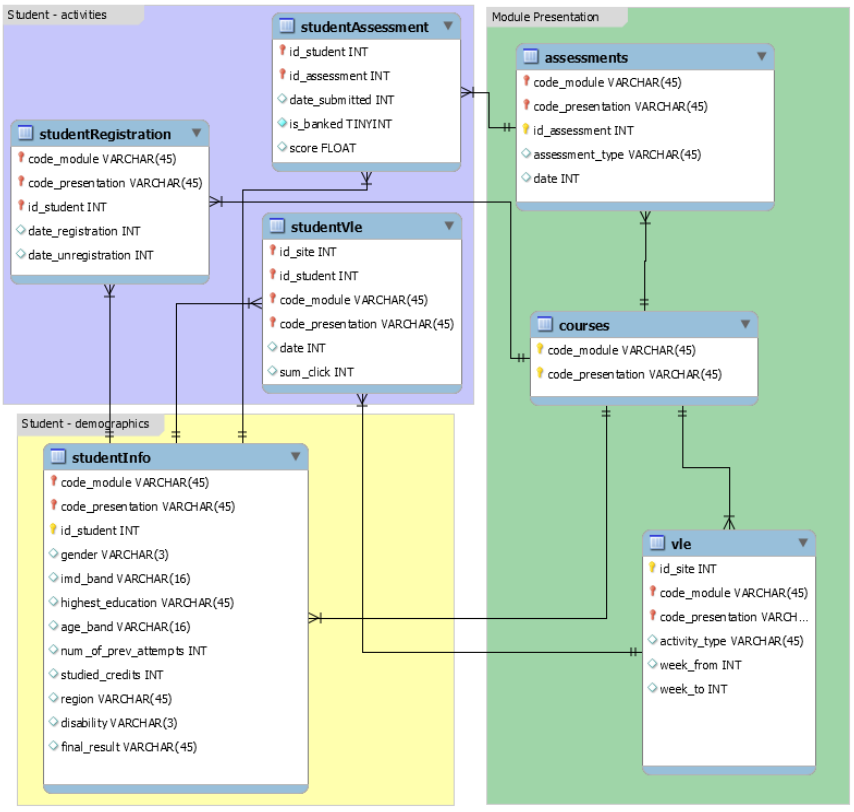

Figure 1. Data Model, Resource:https://analyse.kmi.open.ac.uk/open_dataset Data Wrangling

At first, all data tables mneedes to be merged into one table. Student info table accepted as main table and al other tables are joined into it. However, there were some some issues adding them. First, *student_vle* table's *sum_click* column was suppose to provide one row for each student on a specific day. But, some cases had more than one. Data providers suggested to group these rows and sum the values. Then, the table was ready for joining. Second, there were some students who took the class more than once. Which provided by *num_of_attepmts* in *student_info* table. Therefore, *student_id* was no longer unique id for the table and each row was considered by its own.

Later on, different characteristics of each course had different beahaviours in the data. Such as some courses had exams or some did not. Accordingly, these situations were solved either with imputations or aggregations. Imputations were not derived from data it self, but were provided with the documentation of the data. Like "*If the information about the final exam date is missing, it is at the end of the last presentation week.*"

Finally, the result table *attempt* were prepared and saved. Profile report for this table was prepared with Pandas Profiling[2] which can be found in reports folder.

Exploratory Data Analysis

In this part data is explored to get insights about each attempt's behavioural pattern, scores, interaction with online learning environment, other information like age, gender and so.

While data consisting of 35 columns some were categorical and some were numerical. First, their distributions reviewed.

[2] https://pypi.org/project/pandas-profiling/

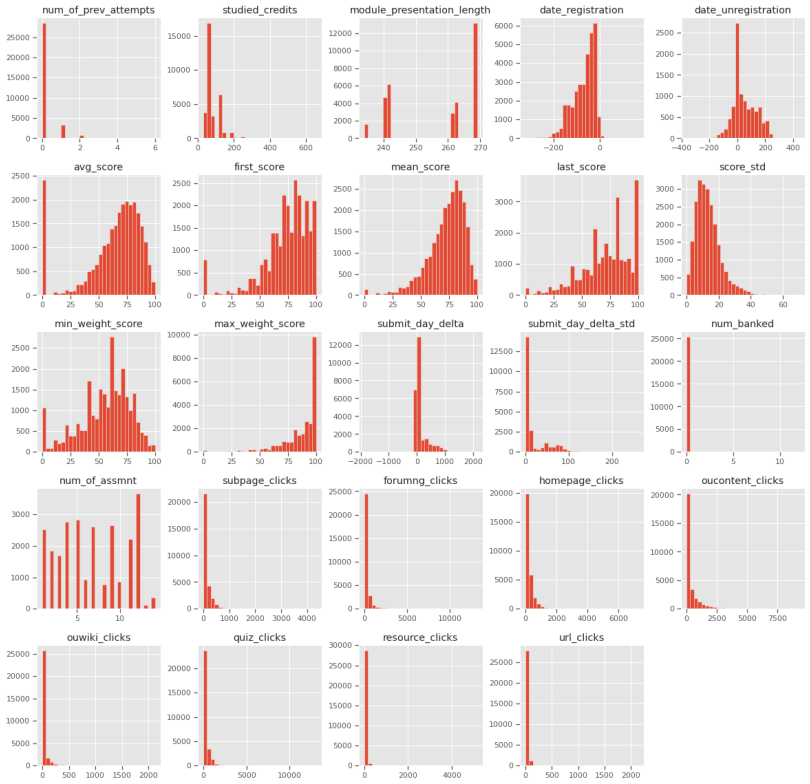

Figure 2. Distributions of numerical features

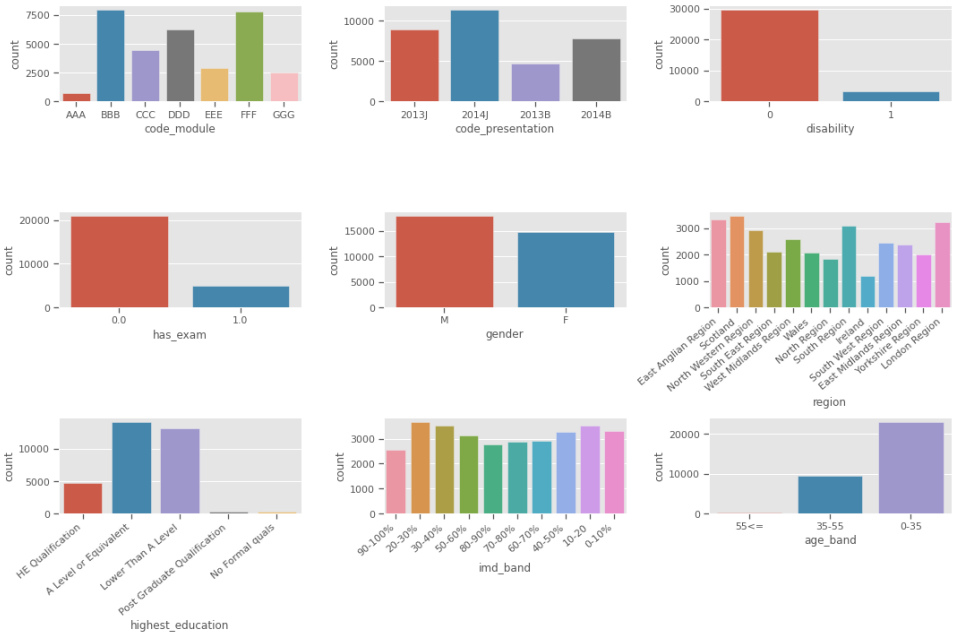

Figure 3. Distributions of categorical features

Later, relations of these features with the target variable *final_result* has compared.

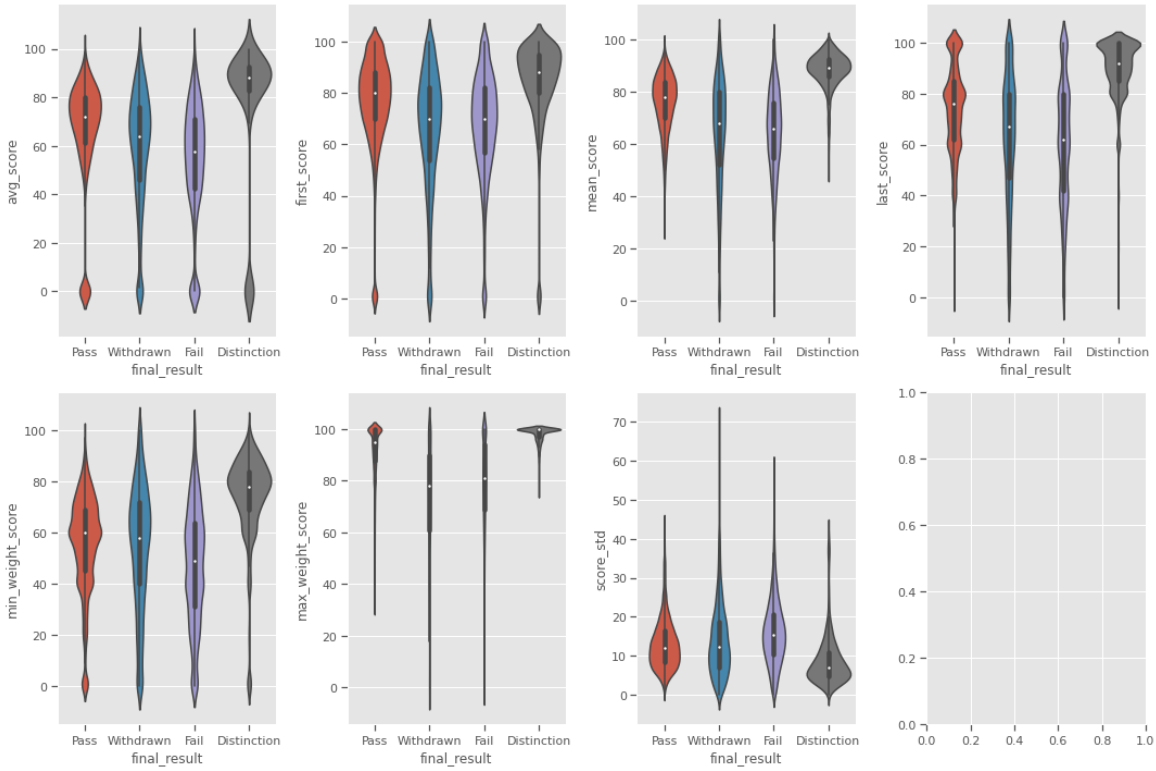

Figure 4. Score related features' relationship with final result

Here there were some attempts with 0 *avg_score* and resulting with distinction. Further investigation revealed that one course had 0 as weights for all the assessments. Which resulted with such situation. And prelacing *avg_score* with *mean_score* just for this course, distributions and relation for both features were similar. Therefore *avg_score* was removed. Relation of *mean_score* with the final result exposed something interesting.

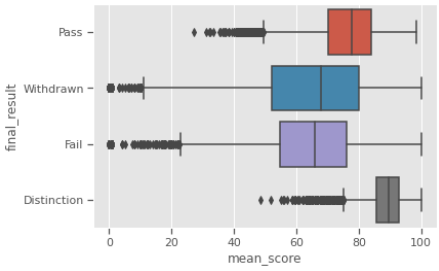

Figure 5. Mean score distributions by the final result

There was a huge overlap of pass and withdrawn or fail. Statistical tests were run to understand if this is an anomaly or not. There were 252 attempts sharing the same means scores but ending with different final results. Having a 32539 total attempts, $H_0 = 0.0077$ $H_1 \neq 0.0077$ and statistical test provided $p = 0.99997$ $CI = (0.007363,\ 0.007373)$. Even if the confidence interval excluded the expected value its very close and such high p_value null hypothesis is denied.

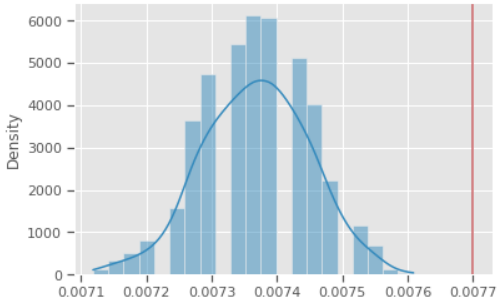

Figure 6. Confidence interval with null hypothesis value

From the categorical features almost none of them showed any specific pattern *region, highest_education* and *imd_band* was worth having a deeper look. While *highest_education* and *imd_band* had nothing special, some regions were not following the same pattern with others.

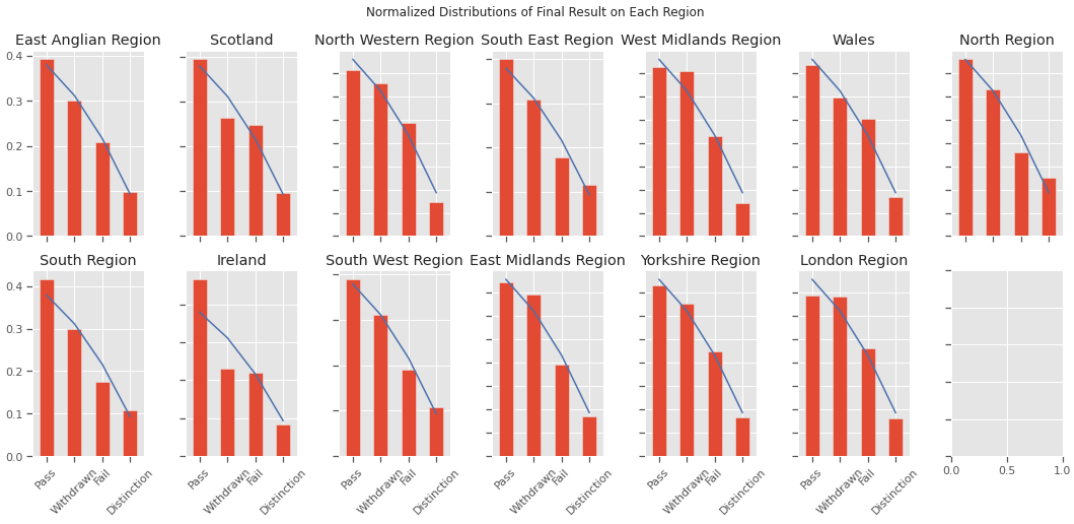

Figure 7. Normalized distributions of final result on each region

Scotland, Ireland and London Region were different than others. Again, statistical tests provided that such occurrences were not region specific.

Modelling

Before modelling missing data analysed to identify MCAR, MAR or MNAR situations. Either related rows/columns removed or imputed. But imputer is fit to train data while transform is applied to train and test data separately.

Prepared data used to develop six different models:

- Logistic regression
- K Nearest Neighbors
- Support Vector Machine
- Random Forest
- Gradient Boost
- Naive Bayes

Target variable was imalanced multiclass feature. So Mathew's Correlation Coefficent and accuracy was decided to evaluate the scores.

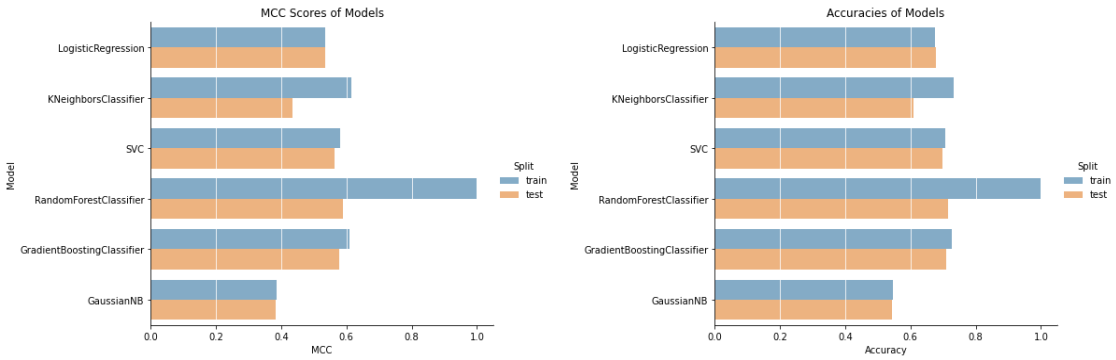

Figure 8. MCC scores and accuracies of the models

The best two models *Random Forest Classifier* and *Gradient Boosting Classifier* chosen for further development. Bayesian Optimization[3] used for fine-tuning and the final results were not significantly different.

	Accuracy	MCC
Random Forest	0.722	0.599
Gradient Boost	0.720	0.596

Eventhough there could be even more improvement of the models they revealed some specific informations. Not only age or imd band being in least importat features, but also homepage clicks being within the top most important ones is interesting.

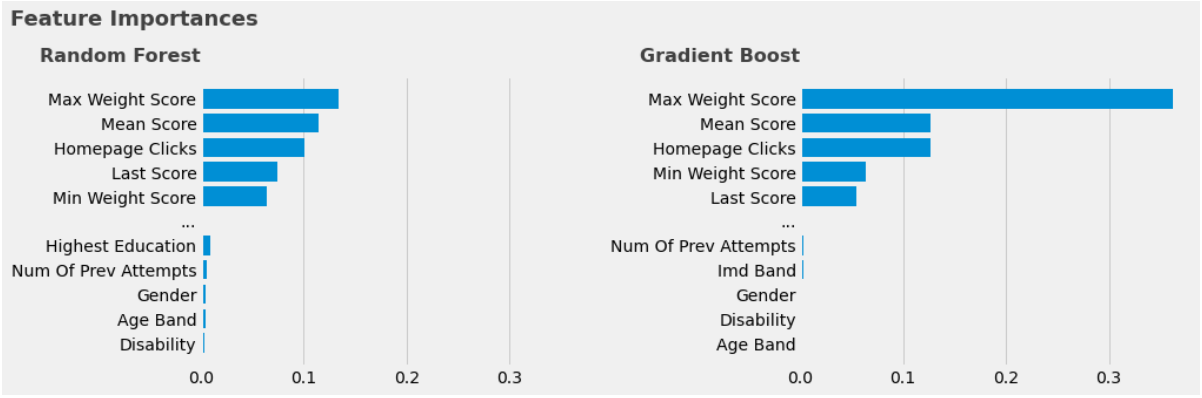

Figure 9. Most and least important features according to tuned models.

[3] https://github.com/fmfn/BayesianOptimization

Conclusion

Both models showed that score and some vle features are the biggest influencers of the final result. And disability, gender and age are being the least. However even after tuning results of the estimators are not significantly changed and the scores are around accuracy: 0.7 and mcc: 0.6 are not really good. At thi point some more features added even at the data collection or some other features can be calculated at EDA - feature engineering.