Open University Learning Analysis

Aishwarya Ramachandrappa
Syracuse University

ABSTRACT

Online education has become very popular and plays a positive impact on learning. Organizations have to consider many factors to make the learning process effective. Virtual Learning Environment (VLE) provided by Open University has several factors which affect the students' performance. If these are identified correctly, better results can be obtained. Both students and faculty can be well informed about the progress based on the analysis which provides an opportunity to excel. Efficient machine learning models are built which predicts the final result.

INTRODUCTION

In recent times, distance learning has become an established part of the education system. It is highly advantageous for full-time workers, military personnel, and nonresidents or individuals in remote regions who are unable to attend classroom lectures. Even regular students can benefit from it to gain subject expertise. But as the interaction between the students and the organization is very less, it becomes hard to analyze the requirements of the students and overcome the shortcomings. It is important to the monitor the student involvement and evaluate their progress over a period.

At Open University, there are about 170,000 students who are a part of several modules. The data collected has enough information of the student to predict their final result. The main idea is to get students involved in modules they would perform better and improve the module content or style if needed. One of the key features that affects the final result directly is the number of clicks. The number of clicks for each student over a module is an important factor as it implies that they have been actively looking into the module and learning from it.

I have performed some initial analysis on the Quasi-identifying attributes [1] to understand the data better. Once the importance of different factors is known, I have built machine learning models to predict the final result. The models are evaluated, and the accuracy graphs are plotted for each of them. For further evaluation, confusion matrix is plotted which gives the True Positive Rate(TPR) and False Positive Rate(FPR).

PRIOR WORK

1. ***Kuzilek, J. et al. Open University Learning Analytics dataset. Sci. Data*** [3]: In this paper, detailed explanation of the dataset along with the data collection method is described. The whole process of data preparation is given. Each table is described along with its attributes. The authors have performed visualization of the data to ensure that all the classes have a good distribution. Representation of demographic data helps us understand the data better and perform classification.

2. ***Predicting student success based on interaction with virtual learning environment*** [5]: In this project, the authors have performed analysis to understand how different features such as region, module, assessment and sum click affects the final result. Further, they use Decision Tree to identify the important attributes and consider the top 10 attributes for classification. Pearson chi-squared is used to evaluate the model.

DATA

The database used is Open University Learning Analytics dataset [2]. It consists of 7 tables with information regarding student activities, demographics and the modules they are enrolled in. This data set consists of details of 32,593 [3] students enrolled in 7 different module at Open University(OU). The data collected is aimed at identifying the featured that affect the students' performance. It is categorized based on 3 main types:

1. ***Demographic***: Gender, disability, age, region, imd-band, highest education.
2. ***Student Activities***: Score, date registered, date unregistered, number of clicks, date submitted.
3. ***Module***: Code module, code presentation, activity, assessment.

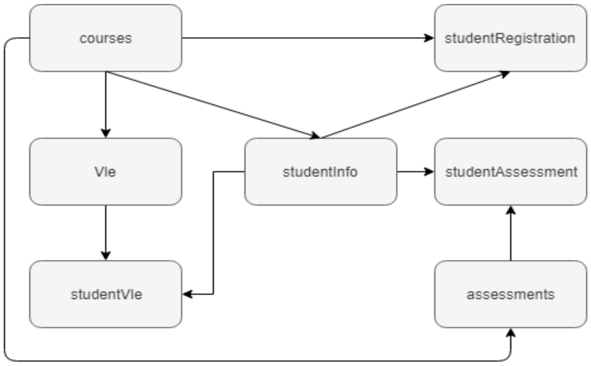


Fig 1. Relationship between the tables.

DATA PREPARATION

Since the data is spread across different tables, it is important to merge all the tables and preprocess it. All the missing values are replaced, and Null/NaN are removed.

After data analysis, all the tables are merged to form the main table which will be used for predictive modelling. Since the data contains columns with text data, it must be converted to categorical data. This is done using Label Encoding. The classes are now replaced by 0 to n-1 labels. This can be used in the models.

DATA ANALYSIS

I have performed some analysis to understand the distribution of the Quasi-identifying attributes. The 6 main attributes identified are gender, disability, age, highest education, region, imd-band. These can be used to analyze its impact on the final result.

1. Gender: Based on the analysis, it can be seen that gender does not play a key role in the result as it is almost the same in all 4 cases.

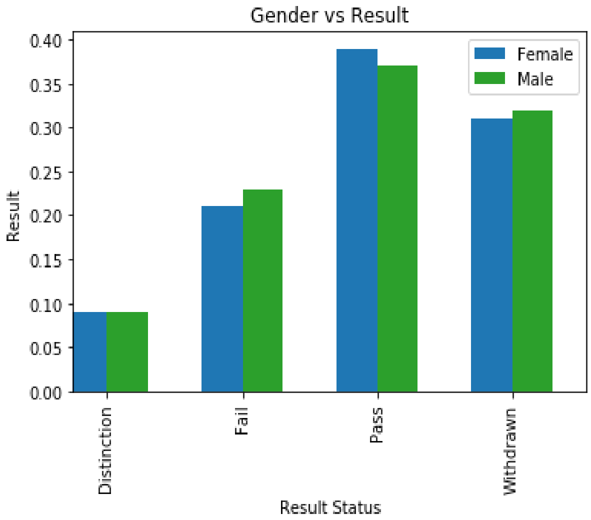


Fig 2. Gender vs Final Result

2. Disability: Based on the analysis, the withdrawal rate is higher for students with disability and the pass percent is low. This can be used as a factor that affects the final result.

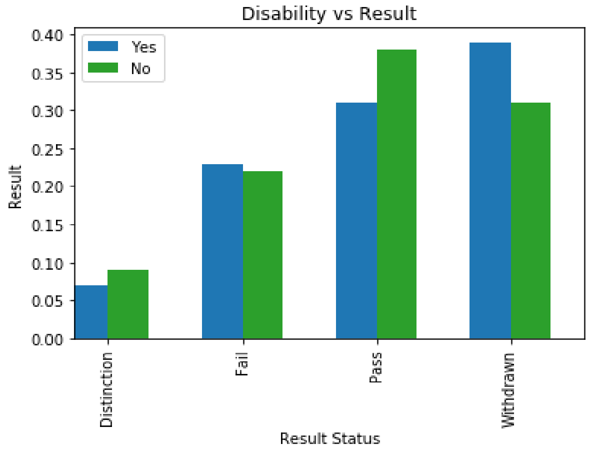
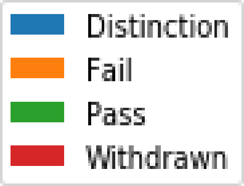


Fig 3. Disability vs Final Result

Legend: Legend for the following bar charts**.**



3. Age: Based on the analysis, withdrawal rate is higher for students under 35. As the age increases, pass percent and distinction also increases. This can be used as a factor that affects the final result.

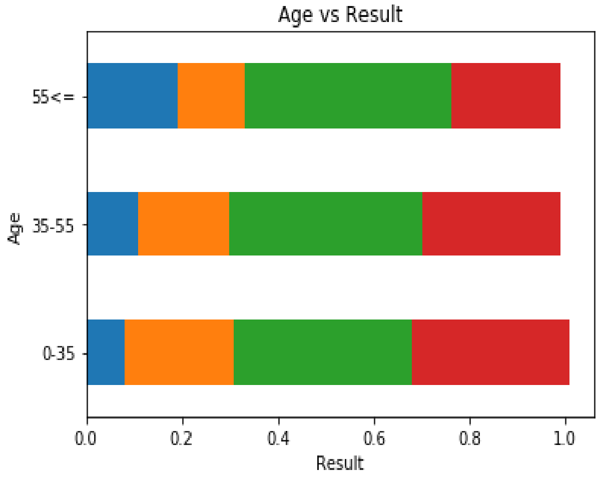


Fig 4. Age vs Result

4.IMD Band: As this IMD rate increases, distinction percentage increases and the fail percentage decreases. This can be used as a factor to predict the final result.

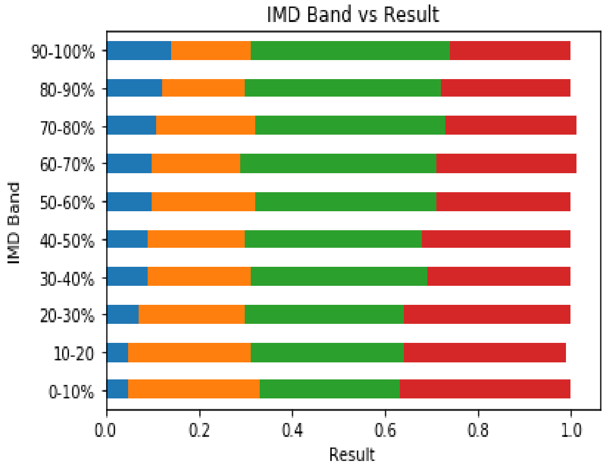


Fig 5. IMD Band vs Result

5.***Education***: Based on the analysis, the withdrawal rate is extremely high if the students don't have any prior formal education. Distinction rate is significantly high in case of post-graduation background.

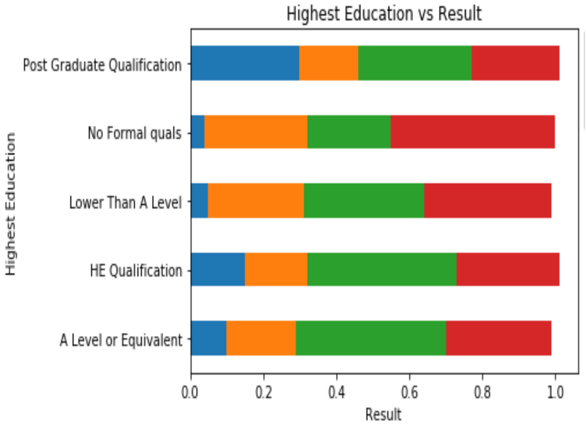


Fig 6. Highest Education vs Result

6.Region: Based on the analysis, the result is almost equally distributed. There is no significant difference between all region and thus it is not a distinguishing factor.

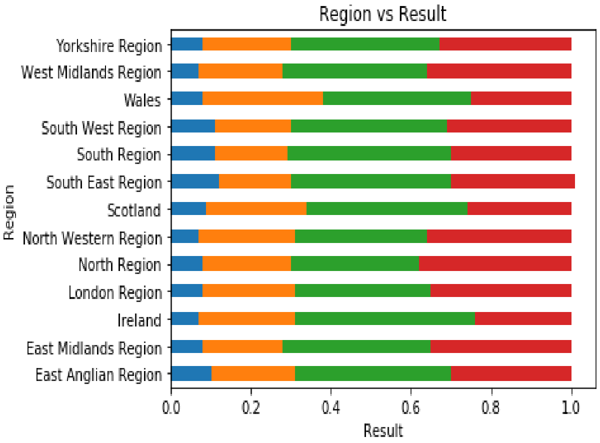


Fig 7. Region vs Result

PREDICTIVE MODELS

Analysis is performed on the merged table. Different models are compared to identify the best model suitable for this dataset. The data is divided in 3:1 ratio for training and testing. The model is trained based on all the factors and then the final result is predicted. All models are built using scikit-learn. [4] In this case, the prediction can be one among the 4 classes – Pass, Fail, Withdrawn or Distinction.

1. ***Decision Tree***: It is a top-down approach of evaluation. In this model, the inner nodes are decisions based on which the data is split, and the leaf nodes are the prediction. Decision trees handles data better when there are huge number of parameters as they consider only one feature at every split. The criteria for split is 'Gini'. The test accuracy obtained is 95.7%.

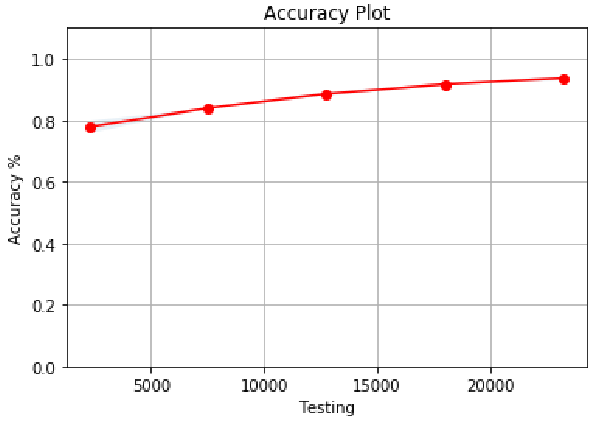


Fig 8. Decision Tree Accuracy

2. ***Gradient Boosting Regression:*** It produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. Thus, gradient boosting combines weak 'learners' into a single strong learner, in an iterative fashion. In the model, the maximum depth is 10 which learns better. The loss function used is 'Huber'. The accuracy obtained is 88.9%.
In the accuracy plot, we can see that there is increase in the accuracy as the depth of the tree increases.

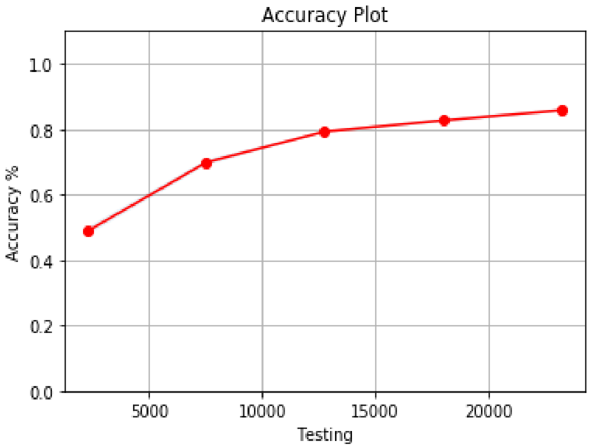


Fig 9. Gradient Boosting Regression Accuracy

3. ***Random Forest:*** A random forest is an ensemble of decision trees created using random variable selection and bootstrap aggregating (bagging). Initially, a group of decision trees are created. For each individual tree, a random sample with replacement of the training data is used for training. Also, at each node of the tree, the split is created by only looking at a random subset of the variables. A commonly used number for each split is the square root of the predictions. It is built using greedy approach selecting the best split points based on purity scores like. The test accuracy obtained is 97.3%.

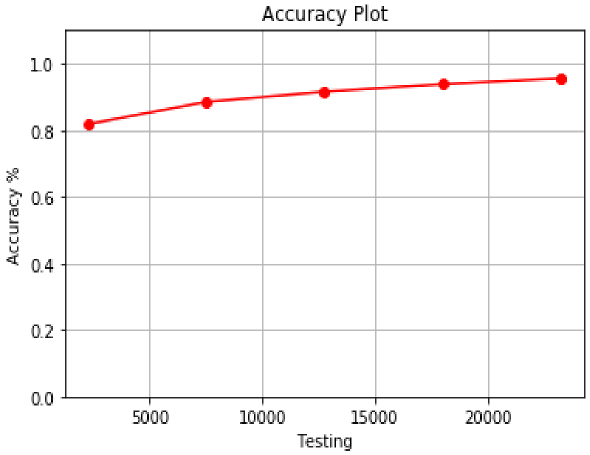


Fig 10. Random Forest Accuracy

EVALUATION

All the 3 models have high accuracy. Random Forest gives the highest accuracy among all. Confusion matrix can be used to identify the TP and FP rates which gives a measure of correctness of the model.

MODEL	ACCURACY
Decision tree	95.7%
Gradient Boosting Regression	88.9%
Random Forest	97.3%

1. ***Decision Tree***: It has high True Positive value indicating that the classes are correctly predicted.

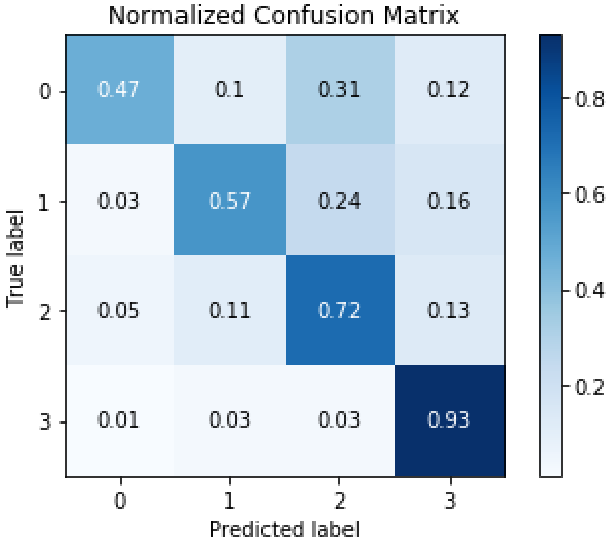


Fig 11. Confusion matrix for Decision Tree

2. ***Gradient Boosting***: It performs well only when the depth is 10. The mislabel (off-diagonal) class is very high indicating that the model is mispredicting one of the classes (0.42). This model is not very efficient as it takes a lot of computations to reach a high accuracy.

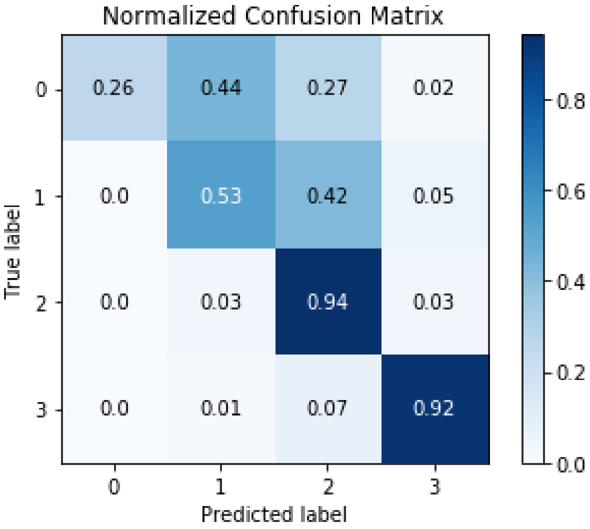


Fig 12. Confusion matrix for Gradient Boosting Regression

3. ***Random Forest***: It has high True Positive value indicating that the classes are predicted correctly. The mislabeled(off-diagonal) is the least in this case and thus proves to be the best method employed.

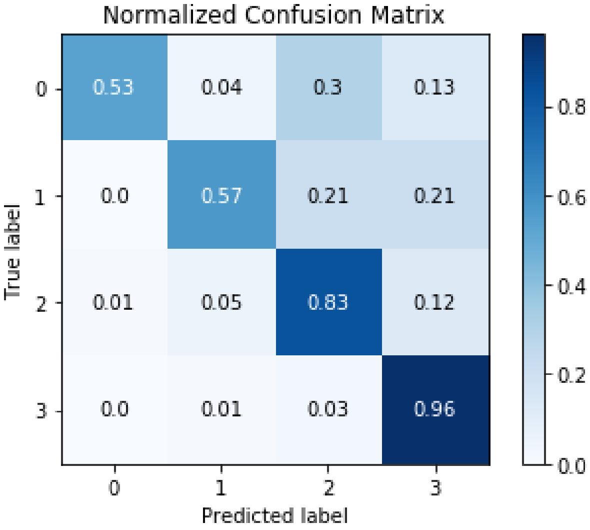


Fig 13. Confusion matrix for Random Forest

CONCLUSION

Analysis and modelling can be used for the betterment of the organization and student.

1. Students' progress should be tracked at regular intervals to ensure positive results.
2. Cases of higher withdrawal rate can be identified and given special interest.
3. Courses can be suggested to students based on their previous performance and course details.

Further, more work can be done to recommend courses to students based on their academic history.

REFERENCES

[1] https://www.nature.com/articles/sdata201717

[2] https://analyse.kmi.open.ac.uk/#open-dataset

[3] Kuzilek, J. et al. Open University Learning Analytics dataset. Sci. Data 4:170171 doi: 10.1038 /sdata.2017.171 (2017).

[4] https://scikit-learn.org/stable/

[5] https://www.lexjansen.com/sesug/2016/EPO-271_Final_PDF.pdf