Predicting Results using ML

I have been investigating the use of two machine learning methods, Random Forest Classification (RFC) and Logistic Regression Classification (LRC) to predict the final result of a student. I will discuss and compare the two methods, in terms of their procedure and performance.

Data Preparation

I acquired a dataset from the OULAD. Grouping the data by category final_result, the number of each grade in the dataset was determined (figure 1). Since logistic regression predicts a binary dependent variable, it was necessary to reduce the four categories to two. Since 'distinction' and 'fail' are the minority, they were eliminated. A heatmap was drawn up from the data to identify missing data (figure 2). A stacked bar chart was then drawn from the category missing data, imd_band, to assess its relevance, fig 3. Since there appears to be a visual trend between IMD band and the final result, rather than eliminating the category completely, entries with missing data were eliminated. 'Pass' and 'Withdrawn' were converted to 1 and 0 respectively. One-hot encoding was used to convert categorical variables such as 'region' to a numerical representation.
To understand better how features affected final grade, further charts were drawn (figure 4). Some categories evidently affect final results more than others.

final_grade	Entries
Distinction	3024
Fail	7052
Pass	12361
Withdrawn	10156

Figure 1

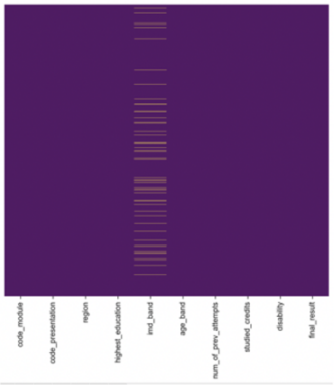


Figure 2

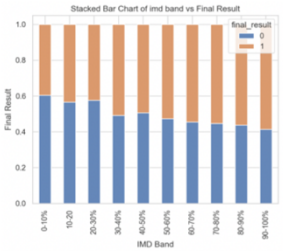


Figure 3

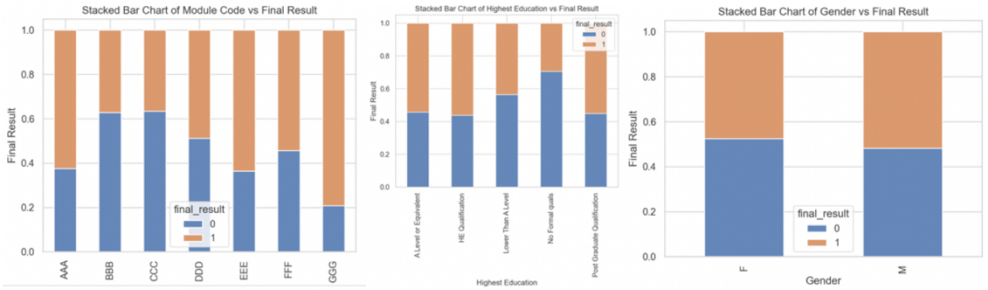


Figure 4

Random Forest Classification (RFC)

RFC involves making a number of decision trees, by determining the category that separates data best, using the Gini Impurity. This method is effective because of its use of bootstrapped data and the fact that a subset of the variables is considered at each stage, creating a variety of trees.

I initially split the data into training and test sets with a 75% and 25% allocation respectively. I instantiated the model with 100 trees, each with a maximum depth of 3 and trained the model on the features and label 'final_result.' I then used the predict method on the test features. Figure 5 shows one of the decision trees generated.

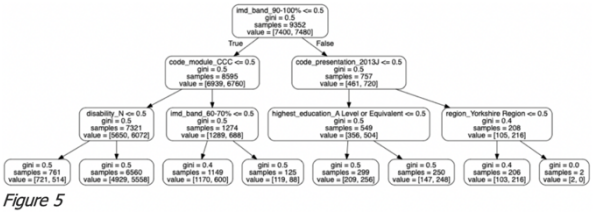


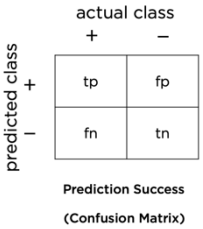
Figure 5

Logistic Regression Classification (LRC)

LRC classifies data through a binary dependent variable by determining the maximum likelihood of a final result value. Data points are transformed into log(odds) values, which in turn are transformed into candidate probabilities using $p = \frac{e^{\log(odds)}}{1+e^{\log(odds)}}$ in order to draw a best fit curve.
Like the random forest classification, the data was split into 25% and 75% for test and training data respectively. The maximum iterations parameter was 100 by default. I fit the model on the training data and test data was used for predictions.

Experimental Procedure

In order to assess the performance of the two methods, I am comparing precision, $\frac{t_p}{t_p+f_p}$, and recall, $\frac{t_p}{t_p+f_n}$, data. The closer to 1 for both measures, the better the performance. I also used confusion matrices and ROC curves, enabling the performance data to be visualised in different ways.
The initial data had 1355 more 'Pass' entries than 'Withdrawn', a percentage difference of 22%. For RFC, despite 0.34 recall for Withdrawn, indicating a prevalence of false negatives, the model is somewhat accurate (fig 6). On the other hand, LRC predicted 100% pass, as shown by its confusion matrix (fig 7).



	precision	recall	f1-score
0	0.66	0.34	0.45
1	0.60	0.85	0.71
accuracy			0.62

Figure 6 - RFC

```
confusion matrix
[[   0 3009]
 [   0 3516]]
```

Figure 7 - LRC

To overcome this, a number of 'Pass' entries were eliminated to achieve an equal 'Pass'/'Withdrawn' split, producing more balanced, accurate data, with RFC slightly more accurate (fig 9), and LRC predicting more Withdrawn grades (fig 10).

final_grade	Entries
Pass	9920
Withdrawn	9920

Figure 8

	precision	recall	f1-score
0	0.66	0.64	0.65
1	0.63	0.65	0.64
accuracy			0.65

Figure 9 - RFC

	precision	recall	f1-score
0	0.52	0.86	0.65
1	0.63	0.23	0.33
accuracy			0.54

Figure 10 - LRC

To investigate the effect of the number of trees on RFC, I ran it with only 10 trees. The decreased precision and recall values (fig 11) indicate more trees improve prediction accuracy. In this case, an increase in trees greater than 100 did not improve performance significantly.

	precision	recall	f1-score
0	0.62	0.60	0.61
1	0.61	0.62	0.61
accuracy			0.61

Figure 11 – RFC with 10 trees

The original dataset includes id_student, behaving as unique identifiers for each data entry which would have no significance on results produced. This was confirmed by comparing prediction results before and after the removal of id_student. Precision and recall increased for both methods.

	precision	recall	f1-score
0	0.63	0.65	0.64
1	0.65	0.63	0.64
accuracy			0.64

Figure 12 – RFC, after removal of id_student

	precision	recall	f1-score
0	0.65	0.65	0.65
1	0.64	0.64	0.64
accuracy			0.64

Figure 13 – LRC, after removal of id_student

Reflecting on the graphs drawn in figure 4, gender did not appear to significantly affect the proportion of each grade achieved. Testing the removal of this category resulted in an increase in precision and recall for both methods. An average of 0.650 was achieved for both precision and recall using RFC and a 0.655 score for precision and recall using LRC.

	precision	recall	f1-score
0	0.66	0.64	0.65
1	0.64	0.66	0.65
accuracy			0.65

Figure 14 – RFC, after further removal of gender

	precision	recall	f1-score
0	0.66	0.64	0.65
1	0.65	0.67	0.66
accuracy			0.65

Figure 15 – LRC, after further removal of gender

I tested the adjustment of the training/testing partition, however no significant difference in performance was noticed.
Under these circumstances, it is clear that one model does not perform better than another.

ROC curves were drawn for both models to help visualise the performance graphically. A ROC curve indicates the ability of the model to distinguish between classes. The greater the area under the curve (AUC), the better the model. The AUC is 0.70 for both models, as shown in fig 10 and 11. Since the AUC is greater than 0.5, the models both seem to produce results that are somewhat reliable.

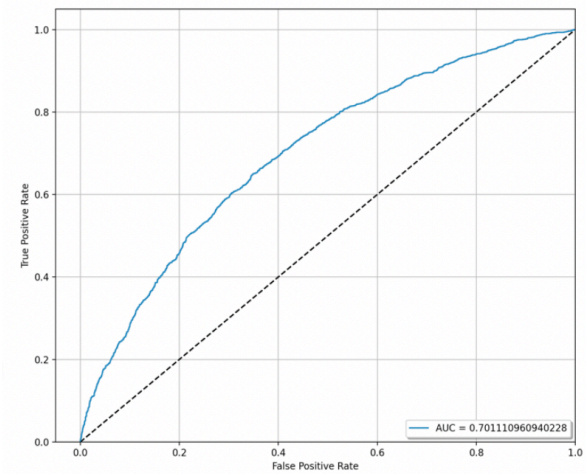


Figure 16 – RFC ROC Curve

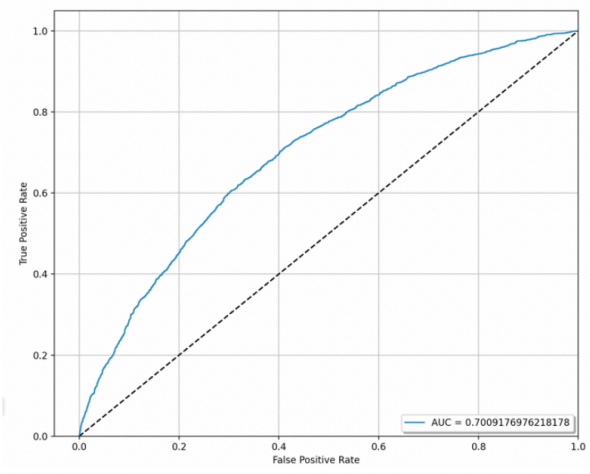


Figure 17 – LRC ROC Curve

Comparison and conclusion

The principles behind these models are completely different – RFC separates data repeatedly by categories that achieve this best, whereas LRC calculates odds and determines the maximum likelihood of a data point. I wanted to see if one approach was superior.
RFC uses bagging, randomly sampling from the dataset with replacement, and ensemble learning, where data from several trees are averaged out. This means small changes in the dataset does not significantly affect predictions. A benefit of this was demonstrated when having an imbalance of data – RFC accuracy decreased by 0.03, whereas LRC was rendered unreliable, with no 'Withdrawn' predictions.
Narrowing features down significantly improved LRC predictions but did not affect RFC predictions as much, already being relatively accurate. This shows RFC is naturally better suited to categorical, multiple choice data because it splits data about category values.
In conclusion, after adjusting parameters and input data appropriately, the overall performances of the two methods were very similar. However, the imbalanced nature of the data and the several multi-variable features made random forest classification more suitable in this case, with more consistent results produced. Instead LRC would be better suited to datasets which are linearly separable and with fewer features.