Machine Learning Coursework

For this assignment I chose Logistic Regression and Decision Tree. After testing logistic regression on a range of features I found that number of clicks on the VLE was a good indicator of the performance. I then combined this with a range of other features in order to gain a more accurate model.

1 Features

In my model, the following features were used

- imd_band
- sum_clicks
- gender
- studied_credits
- num_of_prev_attempts
- highest_education

2 Preprocessing

In the preprocessing for this, I have considered both Withdrawn and Fail to be one type, Fail. I have also merged Pass and Distinction into one type, Pass. Fail is represented by 0 and Pass is represented by 1.

For the IMD band, this was represented in the data given as a range of percentages, I replaced these with the average of the two numbers. The qualifications were given the following ordering

0. No formal quals
1. Lower than A Level
2. A Level or Equivalent
3. HE qualification
4. Post Graduate Qualification

Where there was not complete data for all the features I was using, the data was removed.

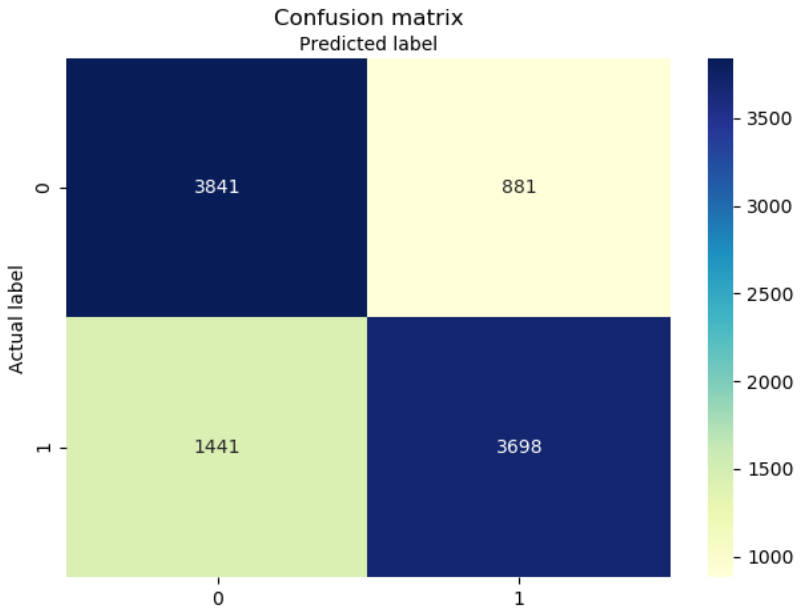
3 Methods

3.1 Logistic Regression

Logistic regression is used as a binary classifier, predicting one of two outcomes, in my case Pass and Fail. This creates a logistic function to predict the outcomes.

In SciKit Learn there are many solvers for logistic regression, they all give similar results, but I noticed the accuracy of the sag and saga solvers was lower for this dataset, and so I chose the solver which gave the best results, liblinear.

With logistic regression the following confusion matrix was generated

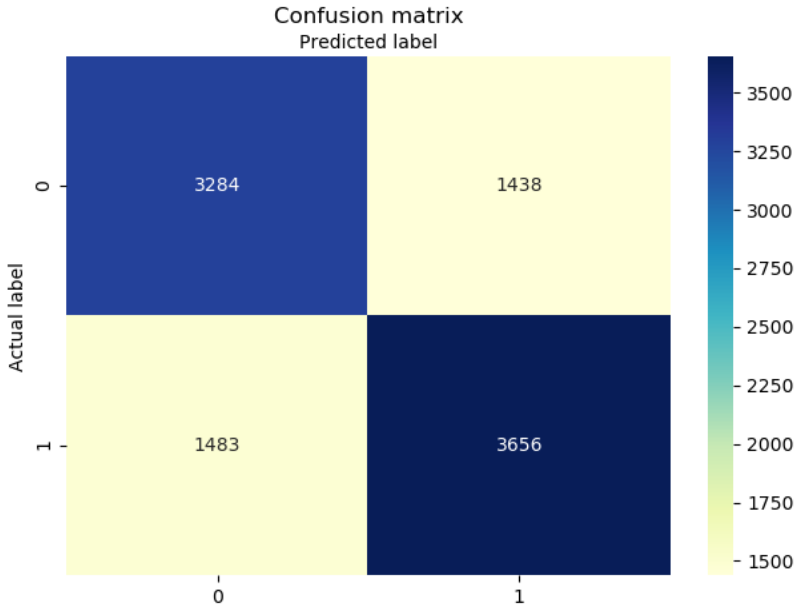


This has an accuracy of 0.76 and a precision of 0.81.

3.2 Decision Tree

Decision tree classifiers work by looking at all the decisions that can be made, an example of a decision in this context is all the qualifications. A tree is then generated which can then be followed when forming predictions.

With the Decision tree classifier the following confusion matrix was generated



This has an accuracy of 0.70 and a precision of 0.72

4 Conclusion

In conclusion on this data set logistic regression has delivered better results, both in terms of accuracy and precision. Number of clicks in the VLE was the best predictor of performance I found, although other features also provide some predictive capacity.