We describe the procedure by which we train both a decision tree, and a random forest model on the OULAD (online learning) dataset. These two models are chosen as they are intrinsically similar, yet the aggregation difference of random forests leads to interesting comparison on performance between the two.

1 Data Preparation

The OULAD data set comes in three categories: student information, assessment data, and VLE data (virtual learning environment). We describe below the transformations that will be applied to the sets.

The student information table holds one entry for each registration of a student onto a specific course (session), with demographic information and their final result, being one of either: distinction, pass, fail or withdrawn, this is the target of the classifiers. Most of the other attributes are categorical, and as such they will need to be encoded or transformed into numeric attributes. Where possible, for categorical numeric data we replace these values with the centre of the range represented, i.e, for age band 0-35 we replace this value with 17.5.

Assessment data is of three types, computer marked (CMA), tutor marked (TMA) and exam. To make this data more sensible, we group the entries by such that there is one to one correlation with the student information table. The mean of the score per assessment type is then taken. Through some exploratory data analysis there are promising correlations between these values and course outcome.

The VLE data is particularly large, with a new entry for each student, activity type and day that they engaged with the course. We look to summarise a students actions over a course on two measures: date, a sense of how many unique days a student engaged with each type of activity; and clicks, a sense of how much interaction a student had with each type of activity, this is inspired by the approach suggested by Jha, Ghergulescu, and Moldovan [1]. To do so we group by activity type over the student and course, then create pivot tables for both unique visit days and total interactions.

To prepare the merged table for the classifiers we pass it through a typical pipeline. Any NA values are replaced by the mean of the axis, except in the case of the VLE data, here rather than fill NAs with the mean of the axis they are replaced by 0. These 0s are what is truly represented in the data, in that if there was no entry for a specific activity type then they never interacted with it. The categorical attributes are one-hot encoded. All numeric values are scaled.

2 Training Evaluation & Tuning

As a starting point we fit both the decision tree and random forest models with the default scikit-learn parameters. We use a 5 fold cross validation strategy to give a good evaluation of their performance (the model is trained on 4 of the 5 folds and tested on the last), avoiding overfitting on the training data when evaluating. We use the metrics of accuracy (correct predictions / total), recall (ratio of true positives to true positives and false negatives), F1 (a weighted average of recall and precision), and ROC AUC, which is the area under the receiver operating characteristics curve. ROC is a plot of the true positive rate against false positive rate, area scores closer to 1 signify better classification.

	Decision Tree	Random Forest
Accuracy	0.810	0.872
Recall	0.865	0.955
F1	0.868	0.916
ROC AUC	0.761	0.916

It is clear to see from this that the random forest classifier outperforms the decision tree. The largest improvement is on ROC AUC, meaning that it has much better class separability, perhaps avoiding some overfitting limitations of the decision tree.

We then proceed to fine tune the models using a randomised hyper-parameter grid search (using ROC AUC score). After tuning we are able to achieve a relatively nominal improvement on the cross validation scores (from 0.916 to 0.919 ROC AUC for the random classifier). We search in the parameter distribution below,

max features	sqrt, log2	
max depth	20, 40, 60, 80, 100, 120, $n-1$	
min samples split	2, 4, 8	
min samples leaf	1, 2, 4	
n estimators	100, 200, 400, 600, 800, 1000, 1200	*forest only*
bootstrap	True, False	*forest only*

This range gives a wide area of sensible values. Tuned parameters can be explored in detail in the classifier. The notable tuned parameter is a max of 800 estimators for the random forest. This high number implies that a wide aggregation is helping to reduce overfitting compared to the decision tree model, giving overall better predictions.

3 Performance & Discussion

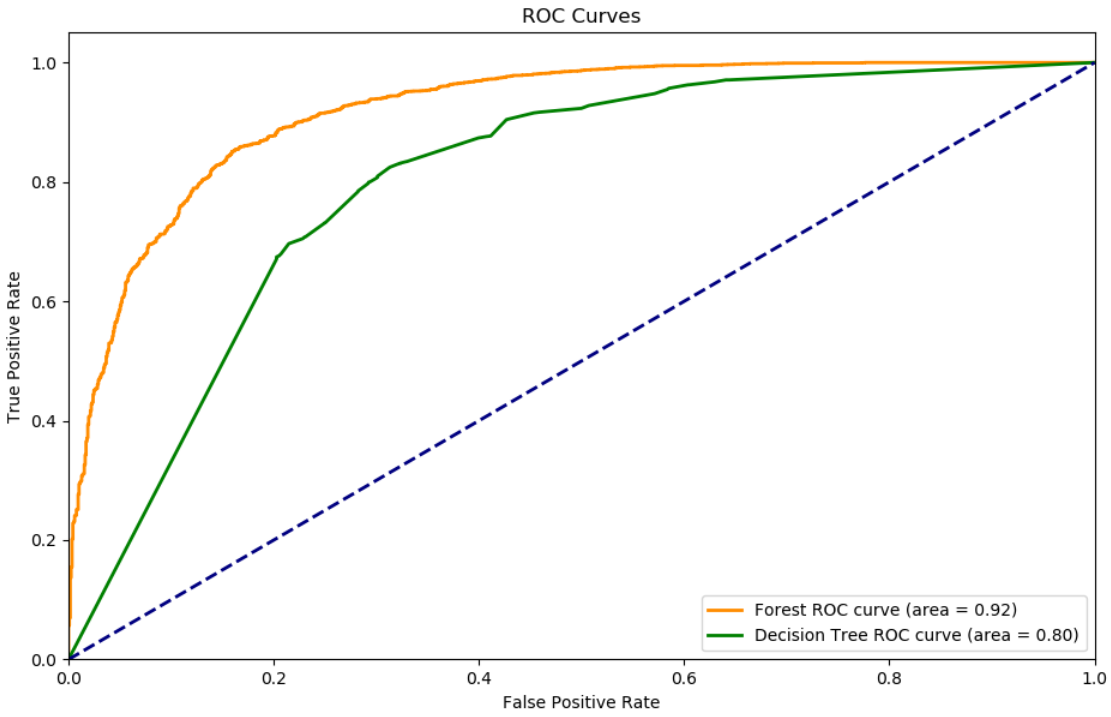
We now run the test data on the tuned models; we start the performance evaluation with confusion matrixes for the two models,

		Decision Tree		Random Forest	
TP	FP	783	386	757	412
FN	TN	506	2549	141	2914

As seen the random forest displays an improvement on the rate of false positives and a marked improvement on the rate of false negatives. Now, the metrics used previously,

	Decision Tree	Random Forest
Accuracy	0.789	0.869
Recall	0.834	0.953
F1	0.851	0.913
ROC AUC	0.801	0.921

The tuned models perform well on the test data, with the random forest model achieving 0.921 ROC AUC. For comparison between the methods find below the ROC curves,



Through this process we can conclude the advantage of the random forest model over decision trees, in the case of unseen testing data. The aggregation aspect limits overfitting, commonly seen in a decision tree model.

This is perhaps most notable in the sharp difference in number of false negatives between the models (506 to 141), which implies that these may be overrepresented in the training data.

The relatively simple model of the random forest allows for fast training and tuning but still offers competitive performance on prediction. On this data set the decision tree is perhaps not sophisticated enough to model the wide range of characteristics.

References

[1] Nikhil Indrashekhar Jha, Ioana Ghergulescu, and Arghir Nicolae Moldovan. "OULAD MOOC dropout and result prediction using ensemble, deep learning and regression techniques". In: *CSEDU 2019 - Proc. 11th Int. Conf. Comput. Support. Educ.* 2 (2019), pp. 154–164. DOI: 10.5220/0007767901540164.