Machine Learning Report

This project aims to predict the final results of students taking online courses as withdrawn, fail, pass or distinction using the OULAD dataset to train two machine learning models.

Data Preparation

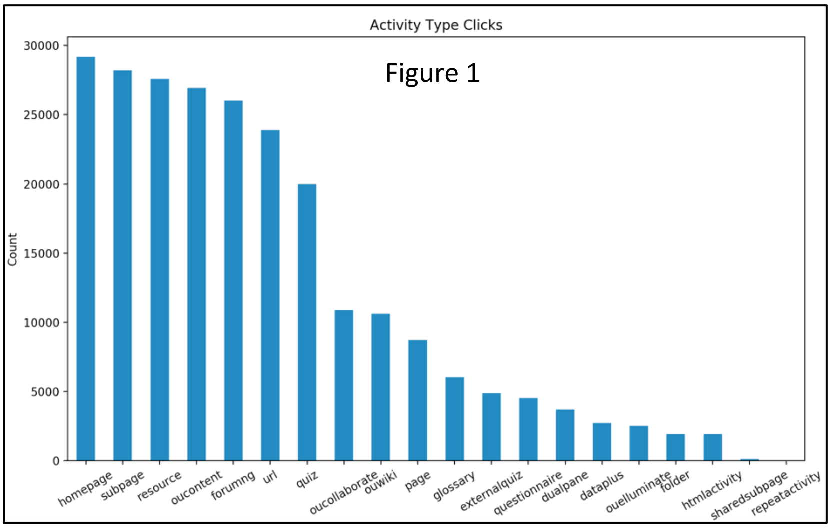
Figure 1

The studentRegistration file was used to calculate whether a student was still registered for a course and this was added to the main dataset.

The mean score for each assessment was calculated. The difference between the student's score and the mean score was then averaged across all assessments the student took then added to the main dataset.

After calculating the mean clicks per activity per student and plotting the activity click count, the eight least clicked activities which had very little data were removed (Figure 1). Any null values for a student's clicks on a remaining activity were set to 0 and added to the main dataset.

Pipeline

A median imputer strategy was found to provide a slight increase in performance over a mean strategy and so was used to replace null values in the dataset as this gave a better accuracy on both models. For numeric data, a standard scaler was used to scale the values. Categoric data was one hot encoded.

Feature Selection

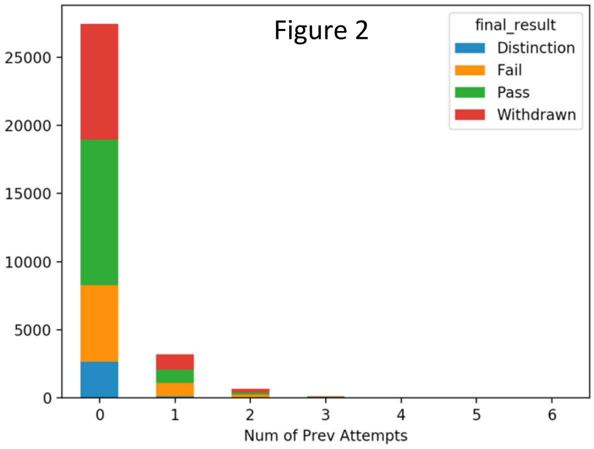
Figure 2

The number of previous attempts, disability, TMA assessed and is_Exam_assessed attributes had very skewed distributions (Figure 2) with most of the data taking the same value and so these attributes were removed from the dataset. After initially running the program, the best random forest estimator was selected and used to plot the relative importance of the data features (Figure 3). Studied_credits, imd band, subpage and ouwiki have the least importance from the model so were also removed from the dataset. An 80-20 train test split was then performed on the dataset.

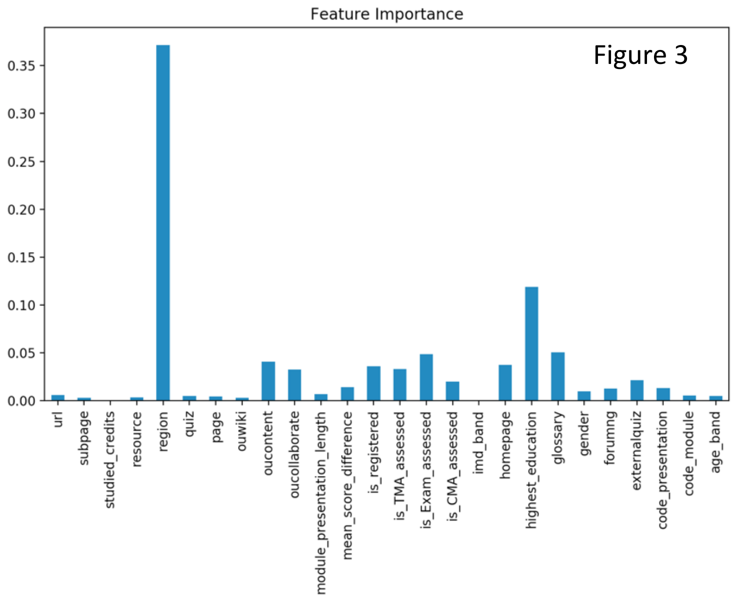

Figure 3

Logistic Regression

Logistic regression estimates the probability a data point belongs to a particular class and then classifies the point using the highest probability. As this is a binary classifier, a one vs rest approach can be used to classify the data into the four classes (pass vs not pass, fail vs not fail etc.). The model estimates the probability by calculating the logistic of the result using the logistic function $\sigma(x) = \frac{1}{1+ e^{-x}}$. This is achieved by maximising the log-likelihood using gradient descent. This model was chosen as it performs well on linearly separable datasets and is less prone to overfitting than other models.

Random Forest Classifier

A random forest classifier is an ensemble of decision trees trained using a bagging classifier. Each tree searches for the best feature from a subset of features when splitting a node, providing greater variation. The relative importance of a particular feature for the prediction can be found by analysing how frequently that feature improves the model. This will allow me to remove irrelevant features from the data after initially training the model. This model was selected as it can be directly applied for multiclass classification rather than a one vs one or one vs rest approach and can classify data with non-linear relationships.

Performance Measurement

F_1 score is the harmonic mean of precision and recall, calculated using the formula $2 \times \frac{(Precision \times Recall)}{(Precision + Recall)}$. As this is a multiclass classification problem, micro averaging will be used to calculate the overall F_1 score. This score can be used to evaluate both of the models.

Log loss is a performance measurement for logistic regression defined by the equation: $J(\theta) = \frac{-1}{m}\sum_{i=1}^{m}[y^{(i)}\log(\hat{p}^{(i)}) + (1-y^{(i)})\log(1-\hat{p}^{(i)})]$ where $\hat{p} = \sigma(x^T\theta)$ (the output probability) and $y = \begin{cases} 0\ if\ \hat{p} < 0.5 \\ 1\ if\ \hat{p} > 0.5 \end{cases}$ (the output class). $e^{-\log\ loss}$ is the probability that the model outputs the right class.

Out of bag score is a performance measurement for random forests and is the proportion of the sampled not in the bootstrapped dataset correctly classified by the random forest.

Parameter Search

Initially a randomised search for twenty iterations is performed on both models to improve the following parameters:

Logistic Regression	Random Forest
C (Inverse Regularisation Strength)	Number of trees in forest
Stopping criteria tolerance	Maximum number of features per tree
Solver for optimisation	Maximum depth of trees in forest

The best estimator from random search is then selected and a grid search is performed to attempt to further locally improve the following parameters by checking values 10% either side of the current best value:

Logistic Regression	Random Forest
C (Inverse Regularisation Strength)	Number of trees in forest

The optimal parameters found for each model were:

Logistic Regression	Random Forest
C=9	n_estimators=648
tol=0.001	max_features=auto
solver=lbfgs	max_depth=18

Model Performance

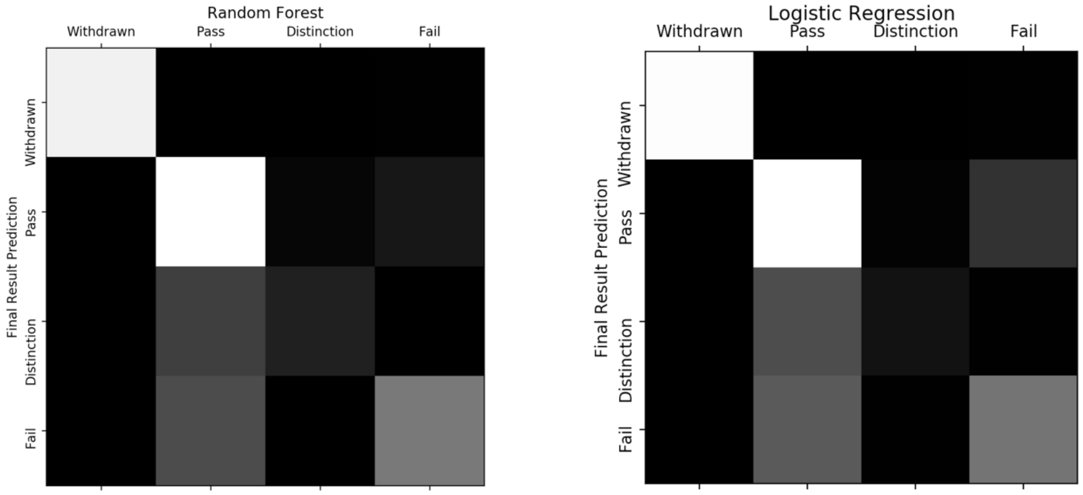

Figure 4

Model Scores		
	Logistic Regression	Random Forest
F_1 Score	0.773	0.819
F_1 Training Data Score	0.779	0.819
Log Loss	0.505	N/A
Average Cross Val Accuracy	0. 769	0. 806
Average Cross Val Training Data Accuracy	0.778	0.816
OOB Score	N/A	0.817

No regularisation was used for logistic regression however the results indicate that neither model experienced overfitting as very similar F_1 scores and cross validation accuracy were achieved by both models on the test and training sets. The OOB score shows that the random forest correctly classified 81.7% of the out of bag rows whilst the log loss of 0.504 indicates that the logistic regression model has a 60.4% chance of correctly predicting a class. Log loss and OOB are not directly comparable; log loss also considers the magnitude of error in prediction.

Conclusion

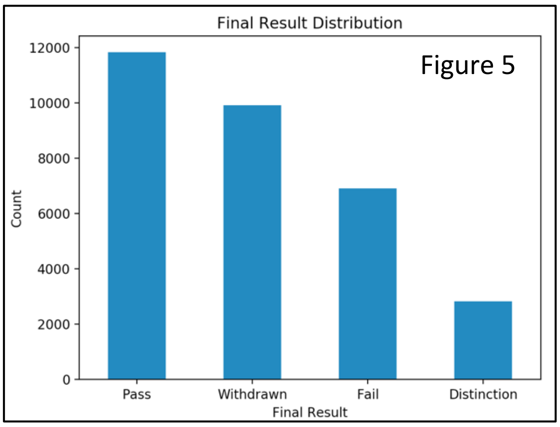

Figure 5

Both the f1 score and cross validation accuracy indicate that the random forest model outperformed logistic regression. This suggests that there may be some non-linear relationships in the data which could not be identified by logistic regression. Further experimentation could be done using non-linear models such as SVMs. As logistic regression is less computationally expensive to train, it is still a viable model as the difference between the scores is small. Random Forest would be the prefered model as it achieved a higher score in every comparable category.

The confusion matrices (Figure 4) show that both models struggled to identify when a student achieved a Distinction result, often misclassifying it as a Pass. Both models also got confused between Fail and Pass but were strong at identifying Pass and Withdraw classes. These classes had far more data values (Figure 5) and the strength of the colour of the diagonals corresponds to the number of data points for each class which suggests that more Distinction examples are required for the models to properly distinguish Distinction from Pass. Another approach would be to use binary classifification and combine Distinction with Pass and Fail with Withdrawn as this would provide sufficient data samples for classification.