PREDICTING STUDENT DROPOUT WITH MACHINE LEARNING MODELS

Stakeholder Report

Prepared by Matt Cocker

Contents

Introduction

Study Group aims to support learners and universities throughout formal education. Study Group needs a model to predict whether students are likely to drop out of full degree programmes at university. That knowledge can then be leveraged to enable Study Group to provide tailored support to students who need it in order to prevent dropout, helping them achieve academic success, whilst also supporting institutions' reputations.

Figure 1: *Project Flow Chart*

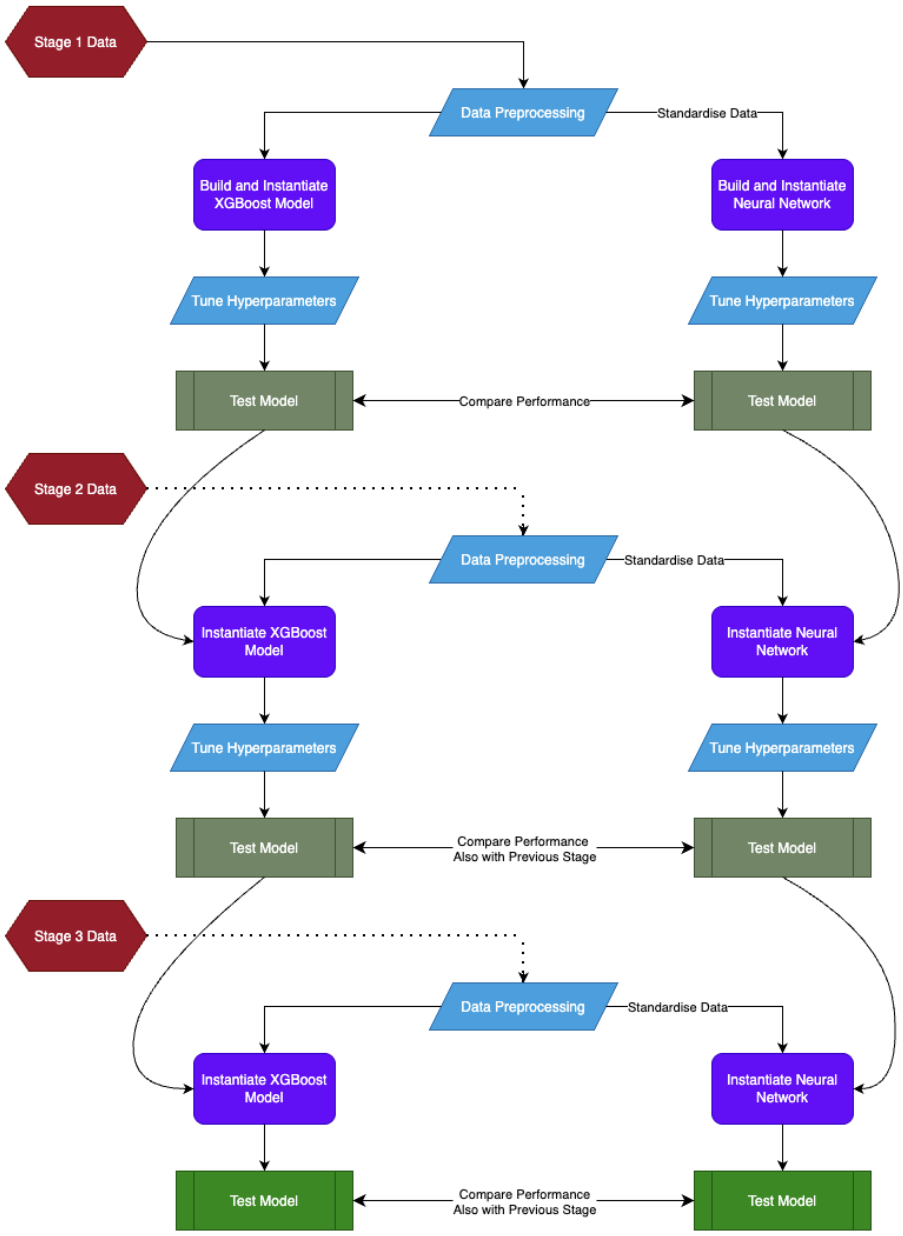


Figure 1 describes the process of creating the model. Data preprocessing ensured missing values were dealt with, and created datasets which could be interpreted by machine learning models. The models were trained and hyperparameter tuning was performed. The models were evaluated on a test dataset to assess their ability to predict student dropout. Models trained and evaluated across the three datasets were compared to reach a conclusion.

Data

The data was split into 3 stages, each containing 25,060 samples. The Stage 1 dataset contained 16 features – basic applicant and course information, Stage 2 contained 18 – the additional two showing student engagement, and Stage 3 contained 21 – the additional 3 showing academic performance. Data preprocessing was conducted to remove irrelevant features, those with a high number of missing values, and those with high cardinality, to ensure a clean and reasonably sized dataset. Ordinal, Binary and One-Hot Encoding were performed to transform the data into a format which could be input into a machine learning model.

The Target variable was visualised to check for imbalance:

Figure 2: *Target Variable Histogram*

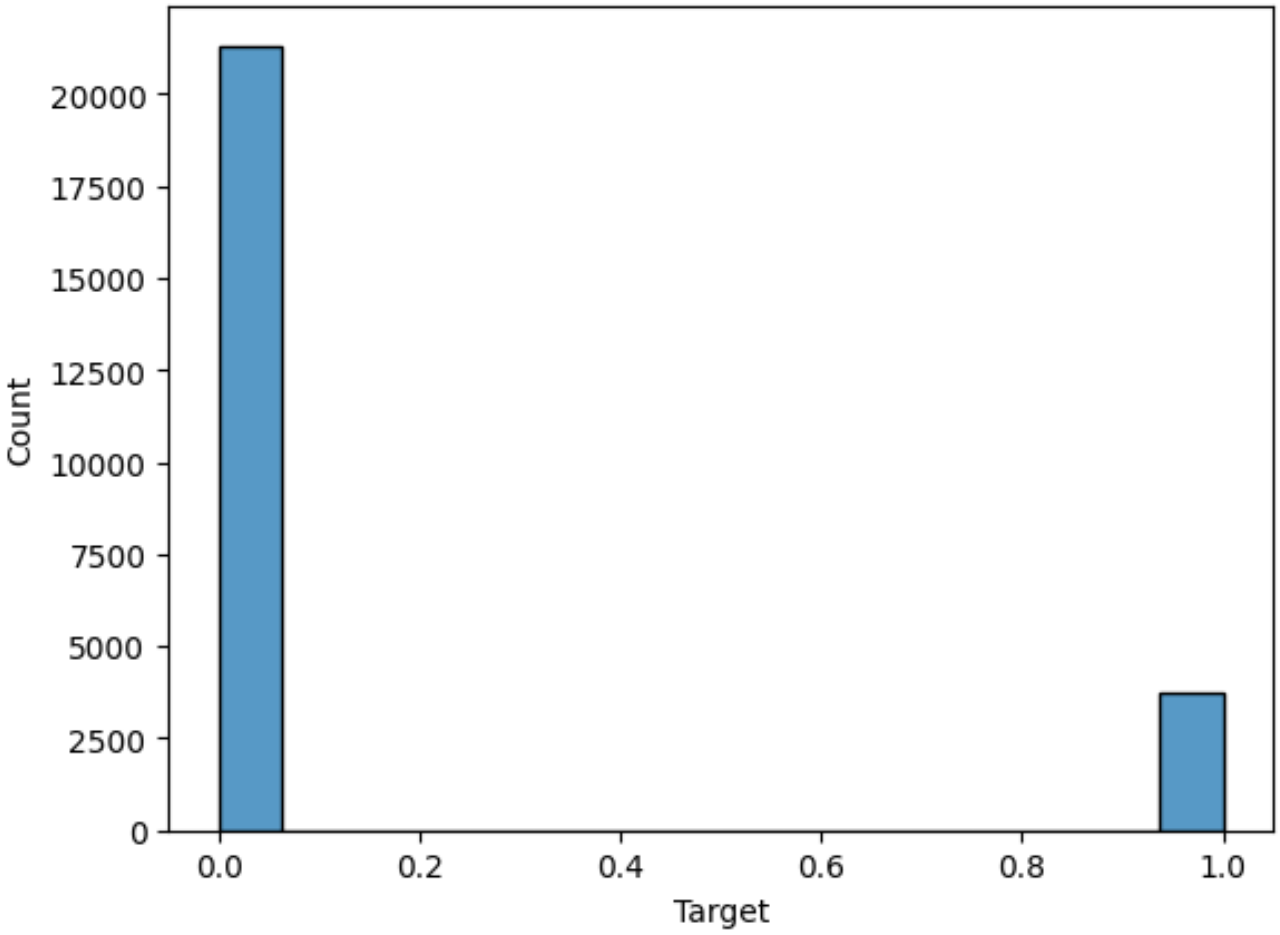


Figure 2 shows clear imbalance in the target variable. However, as this reflects the reality that most students do not drop out, this is not an issue.

Methods

Two different types of models were built – XGBoost and neural network. This is to ensure that a there is not too narrow a focus on a complex solution (neural network) when a simpler one (XGBoost) may suffice.

The hyperparameters of each model were tuned to optimise recall (proportion of true positives correctly predicted), and the tuned models were instantiated on the dataset for the next Stage. This ensures the models are optimised to work on the data available.

- Each model was built, trained and evaluated on the Stage 1 dataset.
- Models were subsequently tuned.
- The models were:
    - Instantiated and trained on the Stage 2 dataset.
    - Compared to assess performance across both datasets.
- Models were then tuned on the Stage 2 dataset.
    - These were compared with models tuned on Stage 1 but trained on Stage 2.
- Models tuned on the Stage 2 dataset were instantiated and trained on the Stage 3 dataset.
    - Results of models trained on Stage 2 and Stage 3 datasets were compared.

XGBoost

XGBoost is an algorithm which builds a series of decision trees, where each new tree tries to improve upon the predictions of the previous ones, with a focus on instances that are harder to predict. It incorporates regularisation and optimisation techniques to achieve high performance and prevent overfitting.

Neural Network

Neural networks take input data and process it through one or more hidden layers. Each layer has a set of neurons, which perform calculations using weights and biases, which are adjusted during training to minimise loss, and an activation function which makes the data non-linear. The data passes through the hidden layers to the output layer, which produces a probability that the sample belongs to one of the two classes.

Results

The results of the models were compared across the three datasets. The Confusion Matrix for the final model at each Stage is displayed below.

The early-stage (Stage 1) models both showed poor performance, even after hyperparameter tuning. This is not surprising, given that generic identifiers such as Gender and Nationality are not good predictors of dropout.

Figure 3: *XGBoost Results (Stage 1)*

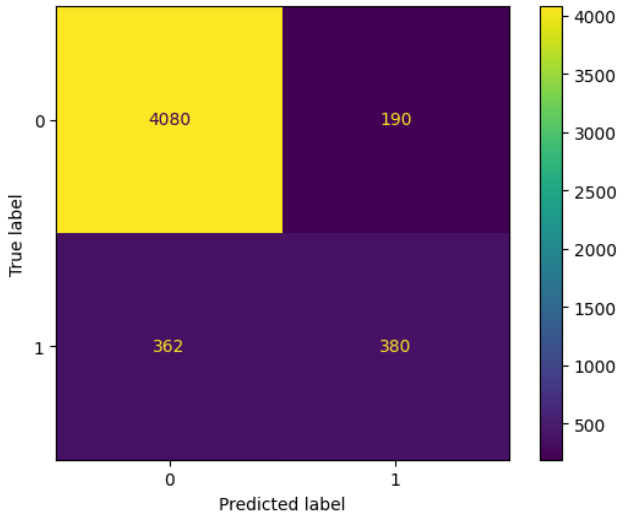
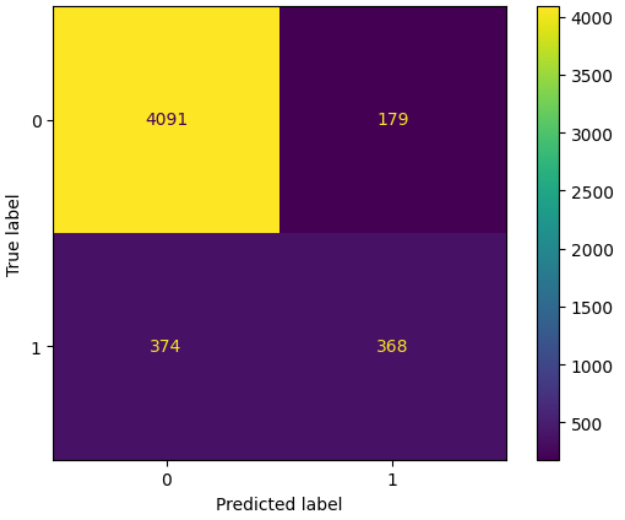


Figure 4: *Neural Network Results (Stage 1)*



Of the 742 students who dropped out (in the Stage 1 Test set), the XGBoost model predicted 51.2% of them, whilst the neural network predicted 49.6% of them. These performances are similar, and show that the data is insufficient to accurately predict dropout.

The mid-course (Stage 2) models showed an improvement in performance, as the features added (Authorised and Unauthorised Absence Count) are indicators of individual student engagement.

Figure 5: *XGBoost Results (Stage 2)*

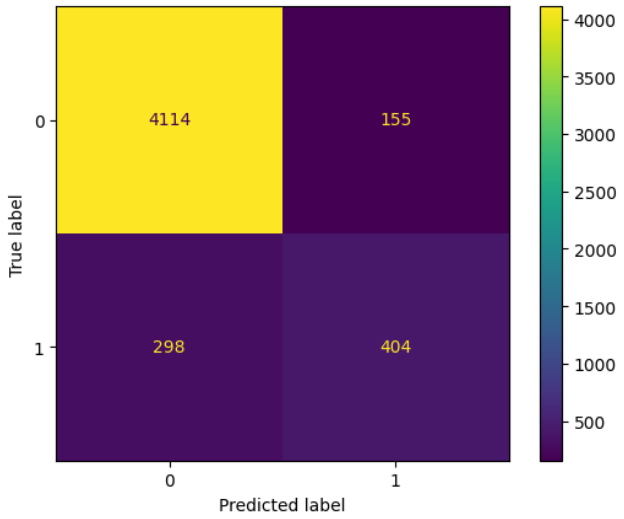
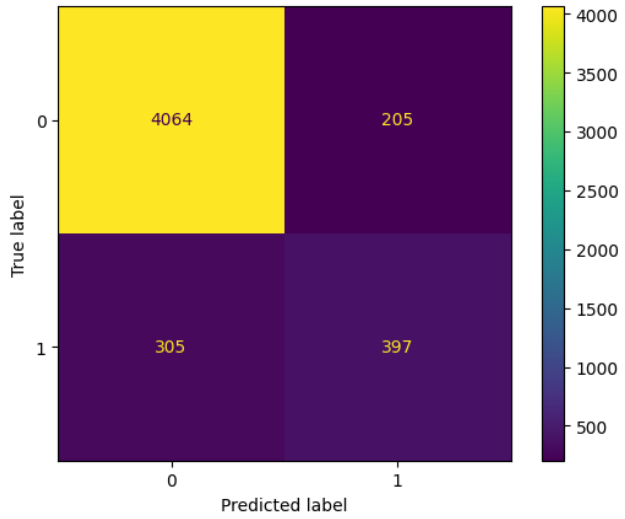


Figure 6: *Neural Network Results (Stage 2)*



Of the 702 students who dropped out (in the Stage 2 test set), the XGBoost model predicted 57.5% of them, whilst the neural network predicted 56.6% of them. These performances are again similar, and showing the data is insufficient to accurately predict dropout.

The late-stage (Stage 3) models added indicators of student performance; students who do not perform well are more likely to drop out.

Figure 7: *XGBoost Results (Stage 3)*

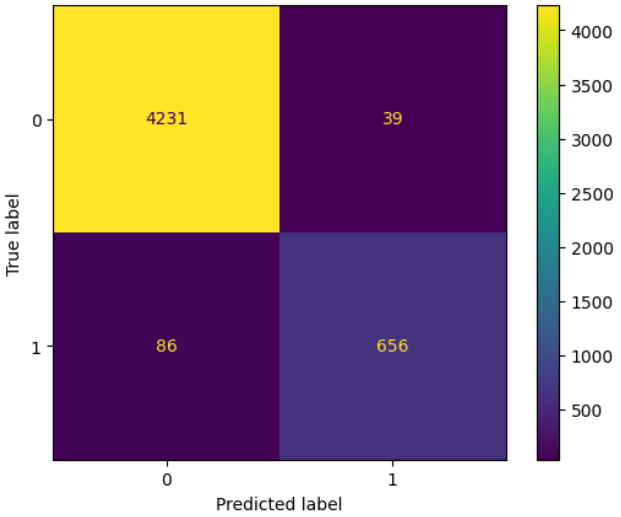
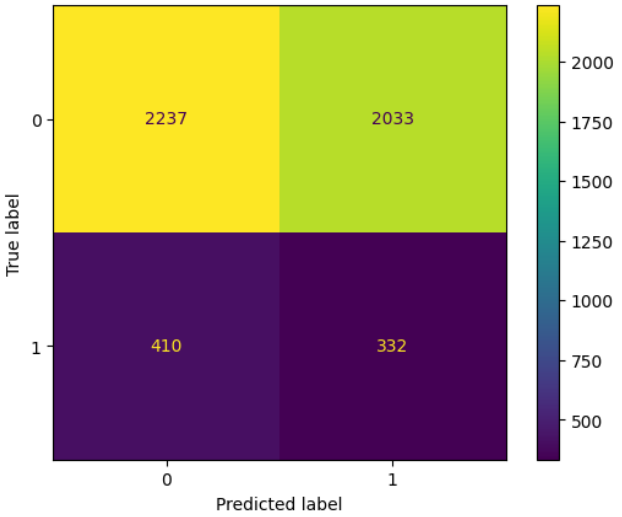


Figure 8: *Neural Network Results (Stage 3)*



Of the 742 students who dropped out (in the Stage 3 test set), the XGBoost model predicted 88.4% of them, whilst the neural network predicted 44.7% of them, and predicted a high number of false positives. These performances are drastically different, suggesting that only the XGBoost model is a good predictor of student dropout rate.

Conclusions & Next Steps

It is clear that the XGBoost model accurately predicts student dropout; this model can be used by Study Group to identify students at risk of dropping out. Hyperparameter tuning of the Stage 3 neural network may improve its performance.

However, predictions can only be made accurately on the late-stage dataset, suggesting that it is difficult to predict student dropout with metrics other than student performance. This limits the time for this prediction to be useful to Study Group. This may be due to an oversaturated dataset, with features adding unnecessary noise to the models.

Next steps would be to use SHAP values to analyse feature importance in XGBoost models, informing feature engineering to generate datasets with less noise and better predictors.

References

Niyogisubizo, J. et. al. (2022)., 'Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization'., Computers and Education: Artificial Intelligence., Volume 3., https://doi.org/10.1016/j.caeai.2022.100066.