Predicting Student Success in Online Courses

Amgoth Pavan Kumar

1. Introduction

The Aim is to build a predictive model that can estimate whether a student will successfully complete a given course based on the factors such as student profile, engagement data and historical data. In this study the following tasks are carried out:

1. Building a classification model to predict whether a student will complete the course or drop out before completion.
2. Analysing feature importance and discuss which factors most strongly influence course completion.
3. Suggesting how the platform could intervene early to help students at risk of dropping out.

2. Data Overview

The data includes three main categories:

- Student Profile Data: Age, gender, major, academic year and region.
- Course Engagement Data: Number of logins per week, videos watched, time spent on the platform, and quiz scores.
- Historical Data: Previous course completion rates, average quiz scores across all courses and the number of courses started but not completed.

The synthetic data generated based on these categories using the Faker library in python. Total 1000 student profiles generated, and classification task carried out to predict the success of a course completion or dropout from the course. As the data is unsupervised no labels are provided to predict the success of a course, to predict the completion status of the course manually a threshold is created.

3. Modelling Approach

Before training the model data need to be processed.

Data Preprocessing: Missing values were handled using imputation methods and categorical variables (gender, region, major) encoded using one-hot encoding, while numerical variables (logins, time spent, quiz scores, etc.,) standardized.

Model Selection: Various classification models are tested, including logistic regression, Random Forest, SVM and XGBoost, with a focus on interpretability and predictive performances.

Feature Engineering: New features derived to capture the interactions between demographic attributes and engagement levels, such as course completion rate and courses not completed. In this process course completed attribute added, which shows whether student completed the course or dropped out indicated with a numerical value (0: dropout 1: complete). A threshold defined to identify the student status on course completion.

Threshold applied to capture the student status on course completion.

```
data['course_completed'] = ( (data['avg_quiz_score'] >= 55) &

(data['avg_score_across_courses'] >= 25) &

 (data['courses_completed'] >= 2) & (data['completion_rate'] >= 0.50)
```

Threshold mainly depends on the average quiz score, average score across courses and completion rate (courses completed / courses started) of the student.

4. Model Evaluation

Depending on the nature of the problem, the following metrics were used to assess model performance: Accuracy, Precision & Recall, F1-Score, ROC-AUC score and Training & Test error to identify whether the model is underfitting or overfitting.

Accuracy:

Table 1. Model and its accuracy

Model	Logistic Regression	Random Forest	SVM	XGBoost
Accuracy(%)	87.0	99.0	91.6	100

Classification Report:

Table 2 Precision, Recall and F1-score of applied models

Model/ Class	**Precision**		**Recall**		**F1-score**	
	Dropout	**Complete**	**Dropout**	**Complete**	**Dropout**	**Complete**
Logistic Regression	0.91	0.69	0.93	0.61	0.92	0.65
Random Forest	0.99	1.0	1.	0.95	0.99	0.97
SVM	0.94	0.80	0.95	0.76	0.95	0.78
XGBoost	1.0	1.0	1.0	1.0	1.0	1.0

Model Results:

XGBoost model achieved the highest accuracy but the feature importance given is less compered to other models. In this case Decision tree model i.e., Random Forest performed good and well with accuracy of 99.0 % and precision greater than 95 % for both classes. Random Forest decrease the variance among the features compared to other models, but the interpretation is less compared to logistic regression. The logistic regression model performance is less compared to RF but the interpretation among the features is high.

In this case RF model shows the higher classification accuracy than other 2 models (i.e., Support vector Machine and Logistic Regression). This happens due to the models works on the decisions of each node of the tree and labels are defined in such manner.

5. Feature importance Analysis

The analysis of feature importance revealed the following key factors influencing course completion:

- Average quiz score across courses
- Quiz Score
- Course completion ratio

The high average score and quiz score during the course are strong predictors of success. The course completion rate is an important feature in predicting whether student dropping or completing the course. If the completion rate is high, the high rate that student complete the course or else drop the course.

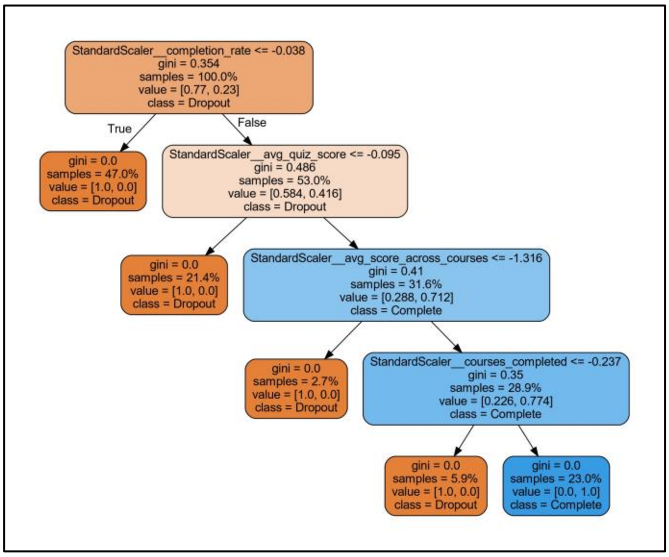

Figure 1. Random Forest Tree

Correlation among the features:

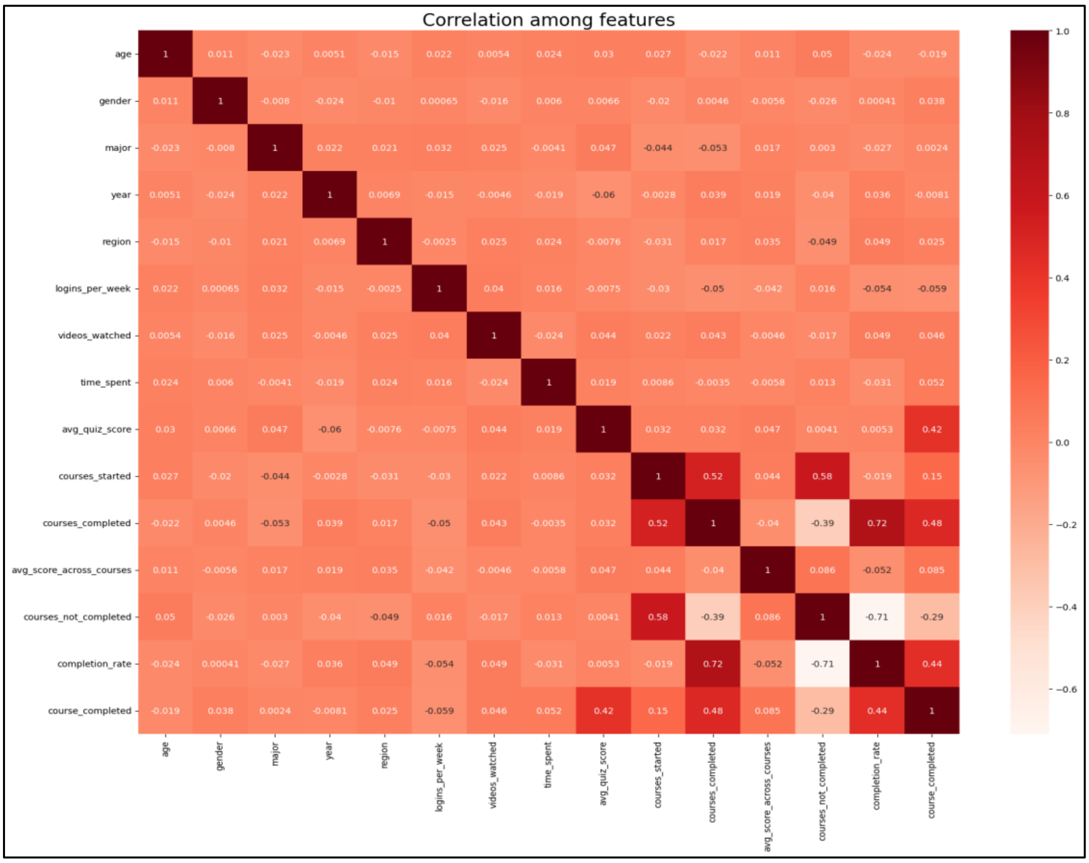

Figure 2. Correlation among the features

The correlation among the features plotted using the confusion matrix. From the matrix there is a negative correlation between the course completion rate and course not started by the student and there is a positive correlation among all features expect with courses not completed feature, it indicated that there is high of a student dropping the course if course not started feature is higher than 5. If the course not started feature is less than or equal to 5, that student will complete the course.

The course completion is also depending on the quiz score, if the student having quiz score higher than 55 and average quiz score higher 25 with completion rate is equal to 0.5 then the chance of student completing the course is high, which can be observed in figure 1 of random forest decision tree.

The pie chart from figure 3 indicates that from this analysis it says that the dropouts are more compared to the course complete.

As the dropping rate is higher, we need to take the precautionary measures and identify the potential features that lead to dropouts. The other features such as time spent on platform, number of logins and videos watched also has impact on these results.

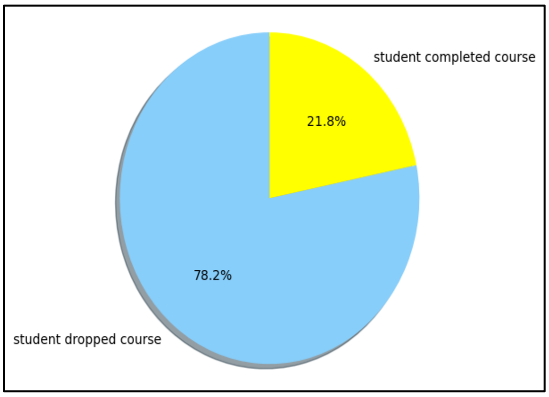

Figure 3. Pie chart of dropout and course complete

6. Recommendations for Early interventions

Based on the models' insights, the following strategies are proposed to help at-risk students:

- ✓ *Personalized alerts*: Sending notifications to students with decreasing login frequencies to encourage more engagement.
- ✓ *Engagement Rewards*: Provide rewards or incentives such as certificates or badges for students who maintain consistent engagement throughout the course.
- ✓ *Customized support*: Offer tutoring or mentorship programs to student with historically lower quiz scores.
- ✓ *Apply Prerequisites*: If the course offered need a prerequisite, start with the basic of the course and make the course more engaging and provide hands on assignments.
- ✓ *Attendance*: Provide the minimum attendance rule to complete the course so that the student who determined can be able to attend and complete the course.

7. Conclusion

The predictive model developed in this analysis effectively identifies students at risk of dropping out from online courses, achieving higher accuracy and providing actionable insights for targeted interventions. The findings from the analysis can significantly help the platform in enhancing the student retention and ensuring higher success rates through data-driven support mechanisms.

Next steps can be done such as model optimization (fine tuning of hyperparameters) and exploring more advanced methods and monitor the impact of suggested interventions on student success. Over time to refine the model and strategies. Collect feedback from the students to better understand their challenges and adjust the predictive model accordingly.

8. Appendix

- **Code:** The code for predictive model is provide in the Jupyter notebook at https://github.com/pavankz/Predicting_Student_Success_in_Online_Courses/blob/main/predict_student_success.ipynb
- **Visualization:** Confusion matrix, ROC curves, errors and feature importances for each model are plotted and included to visualize model performance and feature influence.

Note : Try to open and run the notebook for better visualization of model performance.

> **Thank You** <