Team Progress Report: Developing a Machine Learning Model to Predict Student Dropout Rates or Academic Success

Introduction:

Our team has been working on developing a machine learning model that can predict student dropout rates or academic success based on a variety of factors, such as attendance, grades, and demographic data. The goal of the project is to identify students who are at risk of dropping out or falling behind and provide targeted interventions and support.

Team Members:

Due to unforeseen circumstances, **only one team member** has been able to actively work on the project for the time being.

Work Completed:

Model Development: We started by creating a baseline model using a Random Forest Classifier, which achieved around 75% accuracy. Due to the high dimensionality of the data, we tried tree-based and SVC-based algorithms, but found that tree-based models give better roc_auc scores overall.

Dealing with the Enrolled Class: In our dataset, there is a category or "class" of data labeled "Enrolled." Initially, we thought this class was not relevant for our problem, so we considered removing it or merging it with another class. However, upon analyzing the dataset and the distribution of classes, we decided to keep the "Enrolled" class as it could provide useful information for our model.

We then created two separate models. The first model was trained only on the "Graduated" and "Dropout" classes, which are the classes that we were primarily interested in predicting.

Next, we used the first model to predict the outcomes for the "Enrolled" class in our original dataset. We then replaced the original "Enrolled" labels with the predicted outcomes generated by the first model, effectively creating a new dataset that had predictions for only two classes, but the features of the entire original dataset.

Finally, we used the second model to train on this new dataset that included the predicted outcomes for the "Enrolled" class. By doing so, we were able to improve the accuracy of our final model compared to if we had simply removed the "Enrolled" class from our analysis. The ROC Score improved to 92% from 88%

Counting Encoder: We tested out using a counting encoder to improve the performance of our model. However, we found that it did not help improve the model, so we did not include it in our final solution.

Feature Engineering/Importance: We performed feature engineering by removing highly correlated features, and less useful features from the existing data, but this did not significantly improve the model's performance. We then analyzed feature importance, which helped us identify the most important factors that contribute to student dropout rates or academic success. Results and displayed on GitHub.

Model Deployment: We have deployed our final model on Hugging Face, and anyone can use it to predict student dropout rates or academic success. **This represents an end-to-end solution** that can be used by schools and educational institutions to identify students who are at risk and provide targeted interventions and support.

Next Steps:

Our next steps involve working on the paper and presentation for the project, which will provide an overview of the problem we are trying to solve, the methodology we used, and the results we achieved. We will also continue to explore new features and algorithms that could potentially improve the performance of our model.

Conclusion:

Despite some unforeseen circumstances that have limited the availability of team members, progress has been made in developing a machine learning model to predict student dropout rates or academic success. We have developed a model that achieves higher accuracy than our initial baseline model, and we have deployed it on Hugging Face as an end-to-end solution. Our next steps involve working on the paper and presentation for the project, and continuing to explore new features and algorithms that could potentially improve the performance of our model. We are looking forward to presenting our work and contributing to the field of education research.

Relevant Links:

The Model now lives on HuggingFace:
https://huggingface.co/sulpha/student_academic_success

Relevant codes, notebooks used for testing, training and deployment:
https://github.com/sulphatet/academic-predictor