Predicting Student Dropout Using Machine Learning: A Case Study

Project Abstract

Bryan Chi Fai Pang

Student ID: 501210081

TMU: The Chang School of Continuing Education

CIND 820 Big Data Analytics Project

Dr Tamer Abdou

25 September, 2023

Predicting Student Dropout Using Machine Learning: A Case Study

The issue of university student dropout is a global concern, with dropout rate ranging from 30% in OECD countries to a worrying 50.9% in Costa Rica (Asha, 2020). It has far-reaching economic and societal consequences for all parties involved. For students, dropping out may lead to reduced self-esteem, financial difficulties, and loss of social mobility that a university degree provides. For universities, decrease in student enrollment means reduced tuition revenue, resource allocation challenges and potential damage to institution's reputation. Predictive analytics offers a proactive solution to this challenge by pinpointing students who may be at risk and delivering the necessary assistance to ensure their academic success.

In the article "Predicting Student Dropout and Academic Success," Realinho and his research team introduce a dataset comprising 4423 records (number of students) with 36 features, providing demographic, socioeconomic, macroeconomic, and academic information. They propose the use of four machine learning classification models (Random Forest, XGBoost, LightGBM, and Catboost) for predicting student dropout. However, the article does not provide the actual results of these models in predicting student dropout. Both the dataset and the introductory paper are publicly accessible on the UC Irvine Machine Learning Repository website.

Our research aims to bridge this gap by conducting a comprehensive case study: we actualize the process of building the four classification models suggested by Realinho and extend the scope further with additional models: Artificial Neural Networks (ANN), Support Vector Machines (SVM), k-Nearest Neighbors (KNN), Decision Trees, and Logistic Regression. We select these additional models based on their suitability for predicting student performance, as supported by existing literature. By fitting these classification models using the same dataset, our

research seeks to answer three fundamental questions: Which classification model provides the most accurate classification? Are certain types of machine learning models, such as ensemble models, more effective than others? Does an imbalanced class distribution affect model performance in classification? What are the combined effects of different sampling techniques and classification models? Which features in the dataset contribute the most in predicting student dropout?

To address these questions, we follow a structured approach: Starting with exploratory data analysis using Python and data visualization libraries Matplotlib and Seaborn, we provide an overview of the dataset and then perform data cleaning. We employ different sampling techniques such as SMOTE and ADASYN to mitigate class imbalance, using Imbalanced-learn library. During the process of fitting classification models, we perform feature engineering, selection, and hyperparameter tuning to optimize results, using Scikit Learn and Feature-engine library. We use K-Fold Cross-Validation to evaluate the performance of the models, and present the findings using metrics such as Accuracy, Precision, and F-score. Finally, we evaluate the importance of individual features in predicting student dropout. To conclude our research, we document our challenges and difficulties encountered during the case study and propose enhancements for the prediction model process, as well as potential avenues for future research.

References

Asha, P., Vandana, E., Bhavana, E., & Shankar, K. R. (2020). Predicting University Dropout through Data Analysis. *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*, Tirunelveli, India, 2020, pp. 852-856, doi:10.1109/ICOEI48184.2020.9142882

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success

Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, *7*(11), 146. https://doi.org/10.3390/data7110146

Realinho,Valentim, Vieira Martins,Mónica, Machado,Jorge, and Baptista,Luís. (2021). Predict students' dropout and academic success. UCI Machine Learning Repository. https://doi.org/10.24432/C5MC89

https://archive.ics.uci.edu/dataset/697/predict+students+dropout+and+academic+success