Predicting academic success of undergraduate students

Catherine Meng, Jenson Chang, Jingyuan Wang and Siddarth Subrahmanian

2024-12-06

Table of contents

Summary

In this analysis, we attempt to build a classification model using the k-Nearest Neighbors algorithm to predict student dropout and academic success based on information available at enrollment (including academic path, demographics, and socio-economic factors). Our final classifier performed consistently on unseen test data, achieving a cross-validation training score of 0.71, with a similar test score. Although the model's accuracy is moderate, it performs consistently. Given that the data was collected from a single institution, a larger dataset may be necessary to generalize predictions to other institutions or countries. We believe this model can be a starting point for institution to identify and support students at risk of dropout. However, the model can be developed further by combining academic data with social/economic data to improve the prediction and provide stakeholders with a more comprehensive view on the potential causes of student dropouts. We recommend this improvement because it would enable instutitions to focus their leverage their limited resources for maximum student support.

Introduction

Higher education institutions worldwide face the ongoing challenge of academic dropout and student failure, which affect not only individual students' futures but also the institution's reputation and resources. The ability to predict and anticipate students' potential difficulties is valuable not only for supporting individual students in achieving their goals but also for institutions aiming to implement strategies that support and guide students who may be at risk of academic failure or dropout.

The goal of this analysis is to help reduce academic dropout and failure in higher education by applying machine learning techniques to identify at-risk students early in their academic journey, enabling institutions to implement targeted support strategies.

Methods

Data

The data set is created by Mónica Vieira Martins, Jorge Machado, Luís Baptista and Valentim Realinho at the Instituto Politécnico de Portalegre Realinho et al. (2022). It is sourced from UC Irvine's Machine Learning Repository and can be found here. The data contains demographic, enrollment and academic (1st and 2nd semesters) information on the students. Each row in the data set represents a student record. Using these data, a model would be built to predict the academic outcome of the student. There are 36 columns in total.

Analysis

The Python programming language Python (2021) and the following Python packages were used to perform the analysis: Pandas McKinney et al. (2011), Scikit-learn Kramer and Kramer (2016), Pandera Bantilan (2020) and Altair VanderPlas et al. (2018). The k-Nearest Neighbors (k-NN) algorithm was used to build a classification model to predict whether a student is at risk of dropping out. All variables included in the original data set, with the exception of the Course, Nacionality, Gender, Unemployment rate, Inflation rate, GDP, Previous qualification, Mother qualification Mother occupation, Father qualification, Father occupation columns were used to fit the model. Data was split with 80% being partitioned into the training set and 20% being partitioned into the test set. The hyperparameter K was chosen using 5-fold cross validation. All numeric features were standardized just prior to model fitting. We leave the categorical features as they are because they all have integer data type.

Results & Discussion

To look at whether each of the predictors might be useful to predict the academic outcome, we plotted the distributions of each predictor from the training data set and coloured the distribution by class (Dropout: blue, Enrolled: orange, and Graduate: red).

In Figure 1, although `Unemployment rate`, `Inflation rate` and `GDP` are continous values, they each have less than 10 unique values out of 3000+ rows. This doesn't provide enough range to generalize the problem.

In Figure 2, `Previous qualification`, `Mother qualification`, `Mother occupation`, `Father qualification` and `Father occupation` have cluster patterns but it's unclear what the pattern represents since the ranking of education levels are arbitrary. E.g. "5 - Higher Education - Doctorate" is ranked higher than "1 - Secondary Education" but lower than "10 - 11th Year of Schooling - Not Completed". The source data website provides description on each ranking. `Course` only captures 17 different courses and does not generalize the real world well. `Nactionality` and `Gender` are removed to avoid racial and gender bias

We utilized the k-NN to train the dataset and employed RandomizedSearchCV to fine-tune the hyperparameters. Based on the results, the optimal hyperparameter value is k=19, achieving a best cross-validation score of 0.71. Using this value, we retrained the model and evaluated its performance on the test set, obtaining a final test score of 0.71.

From Table 1, we are only able to predict 68% of the student droputs. Although this will already enable institutions a good baseline to direct their resouces at supporting some of the students, the model could be improved further. In our model, `Unemployment rate`, `Inflation rate` and `GDP` are not used due to the data being skewed. Therefore the model is not using social/economic data data to help predict the outcome of the students. The state of the economy and job market may influence a student's perception of their future post-graduation and and has an impact on the student's academic outcome. Additional social/economic data could be combined with the existing data set to further improve the accuracy of the model. This would also help provide a more comprehensive explanation to instituations on the cause of student dropouts and allow them to target specific issues to address.

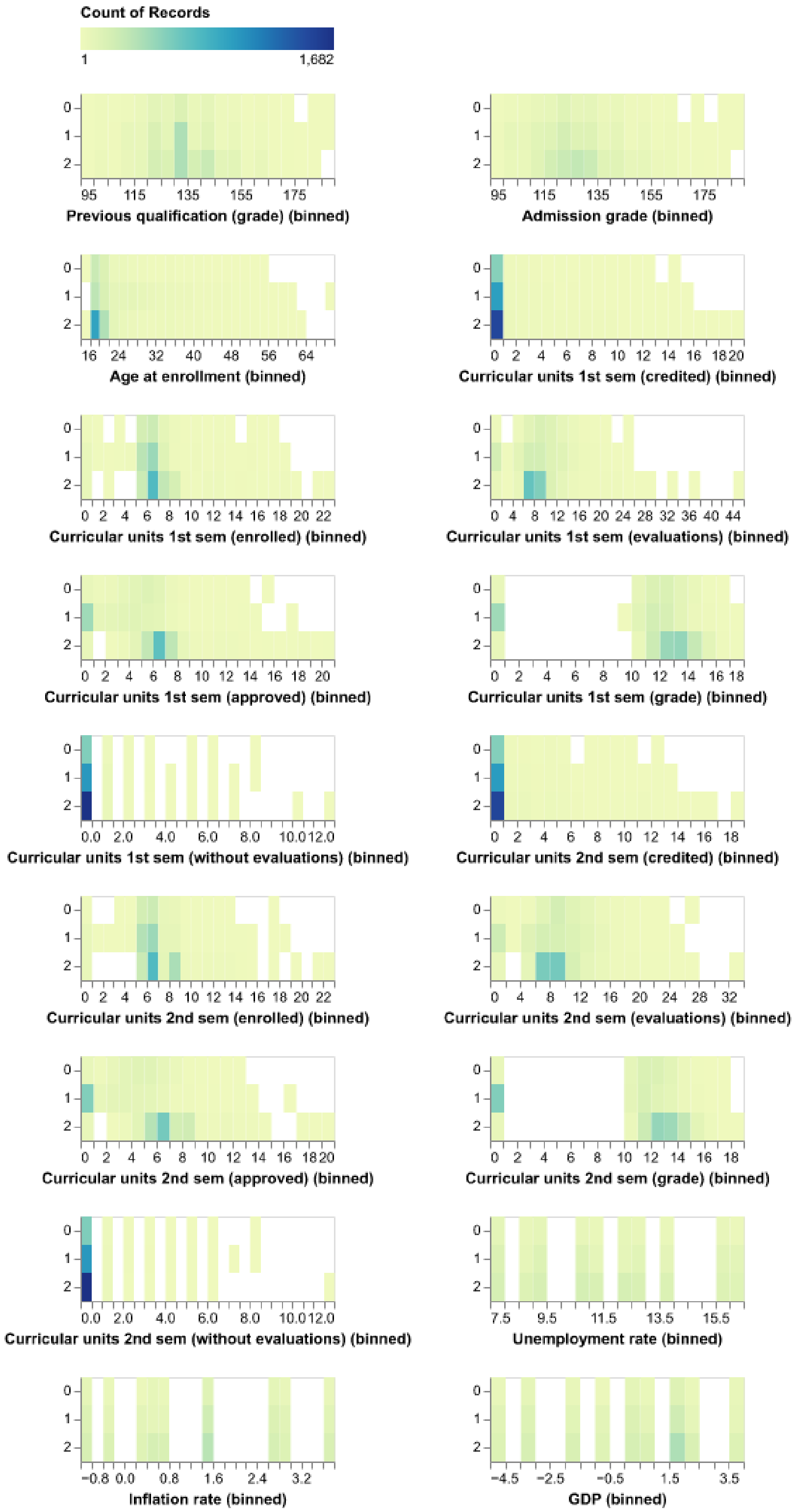


Figure 1: Distribution of Numerical Variable per Academic Outcome

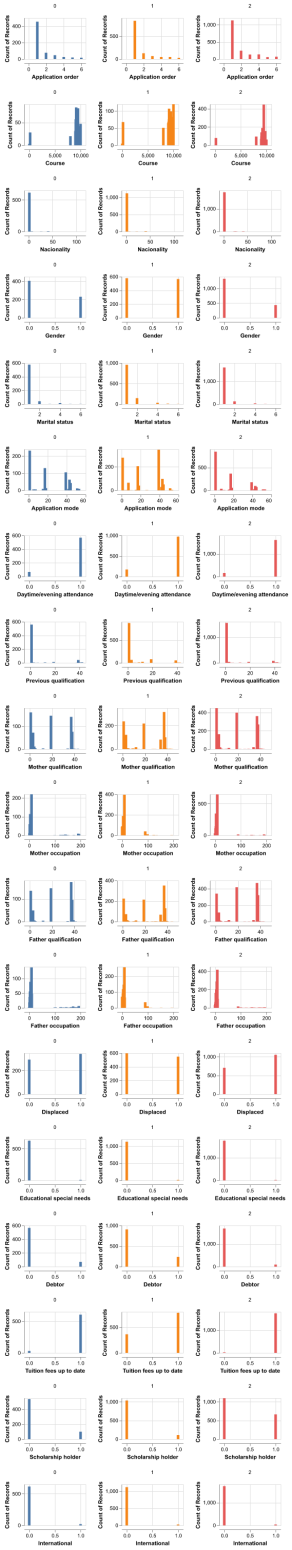
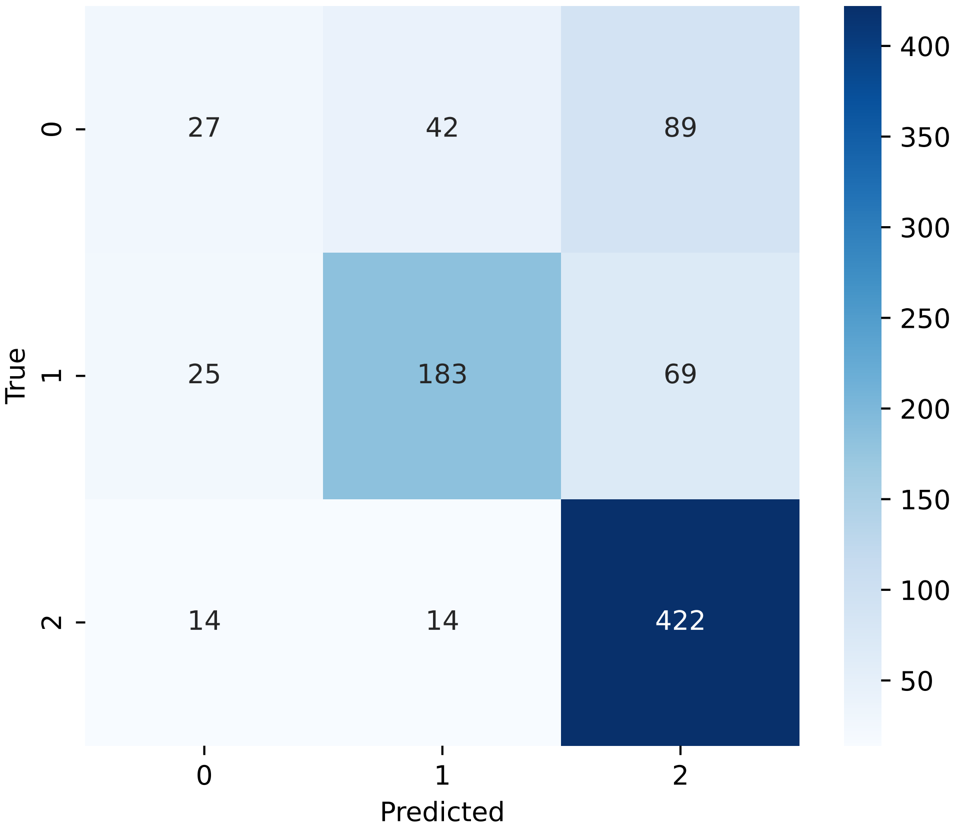


Figure 2: Distribution of Categorical Variable per Academic Outcome

Table 1: Confusion matrix comparing predicted outcomes vs. true outcomes

(a)



True \ Predicted	0	1	2
0	27	42	89
1	25	183	69
2	14	14	422

References

Bantilan, Niels. 2020. "Pandera: Statistical Data Validation of Pandas Dataframes." In *SciPy*, 116–24.

Kramer, Oliver, and Oliver Kramer. 2016. "Scikit-Learn." *Machine Learning for Evolution Strategies*, 45–53.

McKinney, Wes et al. 2011. "Pandas: A Foundational Python Library for Data Analysis and Statistics." *Python for High Performance and Scientific Computing* 14 (9): 1–9.

Python, Why. 2021. "Python." *Python Releases for Windows* 24.

Realinho, Valentim, Jorge Machado, Luís Baptista, and Mónica V Martins. 2022. "Predicting Student Dropout and Academic Success." *Data* 7 (11): 146.

VanderPlas, Jacob, Brian Granger, Jeffrey Heer, Dominik Moritz, Kanit Wongsuphasawat, Arvind Satyanarayan, Eitan Lees, Ilia Timofeev, Ben Welsh, and Scott Sievert. 2018. "Altair: Interactive Statistical Visualizations for Python." *Journal of Open Source Software* 3 (32): 1057.